Germline mutation in rare disease

Joanna Kaplanis

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Downing College September 2020

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except where specified in the text. This dissertation does not exceed the prescribed world limit set out by the Degree Committee for the Faculty of Biology.

Joanna Kaplanis
September 2020

Acknowledgements

I would like to start by thanking my supervisor, Matt Hurles. I feel very thankful to have had the opportunity to work on some truly exciting projects in the last few years. I have learnt as much about the scientific process as science itself from you, especially in your ability to balance your remarkable scientific intuition with evidence and trust in the data. Your ability to notice when the slightest detail in a figure is not quite right has set me back on the right track several times. Aside from your scientific support, I am especially grateful for your kindness and encouragement when I made the decision to have a child in the middle of my PhD.

I would like to thank everyone involved in the DDD project for your work in creating such an important resource and for all of the scientific input from meeting discussions. I am especially grateful to all of the children and their families who have participated in DDD as well as those who have contributed to other cohorts included in this work. Thank you to the Wellcome Trust for funding this work and to Annabel Smith, Christina Hedberg-Delouka and the committee of graduate studies for keeping the Sanger PhD programme running smoothly.

Thank you to the members of Team 29, past and present, for engaging scientific discussions, fun lunchtime chats and all of that delicious cake. I am especially grateful to Kaitlin who was a wonderful scientific teammate and always made time for my questions; I will miss our endless to-do lists. A special thanks also to Eugene, for always being willing to help and advise. Thank you to Carol Dunbar for handing all of the logistics of Team 29. Thank you also to everyone beyond Team 29 who has made the Sanger a fun place to be these last 5 years especially to Eva, Sophie, Alex, Fernando, the members of the now defunct breakfast club and the free biscuits at HumGen tea.

Lastly, I want to thank my family for their unwavering support in the last few years. To Paddy, for moving to Cambridge with me and for your enthusiasm and belief in me. To Theodora, my other thesis chapter, for your smiles at the end of the day and helping me keep everything in perspective. To my brother, who has always inspired me and managed to start and finish a PhD within mine. Finally to my parents, who never fail to pick up the phone, for their love, encouragement and patience. I would never have been able to do this without you, thank you for all of the opportunities you have given me.

Abstract

Germline mutation is the ultimate source of evolutionary change and disease-causing variants. Understanding the rates and patterns of human mutation can help us learn about their molecular origins, uncover our evolutionary history and improve our ability to identify the genetic causes of human disease. With the advent of exome and genome data sets of parent-offspring trios there is an unprecedented opportunity to characterise mutations at an individual level and to harness the increasing sample sizes to identify disease-causing mutations. The goal of this thesis is to understand sources of variation in germline mutation and the contribution of these mutations to rare developmental disorders. These sources of variation encompass types of mutations that have been previously underrepresented in genetic research as well as individual mutation rates and spectra across individuals and parental origin. These analyses fall into three distinct projects.

My first project in this dissertation focuses on the mutational origins and pathogenic impact of multi-nucleotide variants (MNVs). These are variants that fall within 20 base pairs of each other and are frequently misannotated in variant-calling pipelines. Using data from the Deciphering Developmental Disorders (DDD) study, I explore the pathogenicity of this type of variant and found that MNVs in protein-coding sequences can be more pathogenic than a single nucleotide variant even when the MNV falls within a single codon. I also estimate the MNV mutation rate, explore the mutational spectra of these variants and describe the contribution of *de novo* MNVs to severe developmental disorders.

The next project focuses on identifying and characterising germline hypermutators. Using sequencing data from the DDD and 100,000 Genomes Project datasets across ~20,000 parent-offspring trios, I identified fifteen children with an unusually large number of *de novo* mutations. Eight of these appear to be due to a paternal hypermutator. I describe analyses to try and identify a genetic cause for this hypermutation. For two of the individuals, I found rare homozygous paternal variants that fell into two different DNA repair genes and are the likely cause. I also explore whether variants in DNA repair genes more generally impact germline mutation rates. First by examining a well characterised cancer somatic mutator gene and second by using a broader approach across all DNA repair genes. Using the large resource of DNMs called in the 100,0000 Genomes Project dataset, I also estimate what fraction of variance in germline mutation rate can be explained by hypermutation as well as by parental age.

In my final project, I describe analyses of *de novo* mutations in a cohort of individuals with developmental disorders (DDs). *De novo* mutations are a major cause of DDs however known genes only account for a minority of the observed excess of these mutations. Here I develop a statistical framework and apply this on *de novo* mutations from ~31,000 exome sequenced parent offspring trios from the DDD study pooled with trios from GeneDx, a US-based genetic diagnostic company, and trios from Radboud University Medical Center (RUMC). I identify 28 genes that were not previously robustly associated with DDs and explore how these genes differ from those that were previously known. I also develop a model-based approach to explore the likely properties of currently undiscovered genes which can inform future directions in the field.

Collectively, these results reveal important insights into sources of variation in germline mutation rates as well as in mutation type. This can inform how germline mutations arise and further improve our ability to assess their contribution to rare genetic disease.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation

Germline mutation is the ultimate source of evolutionary change and disease-causing variants. Understanding the rates and patterns of human mutation can help us learn about their molecular origins, uncover our evolutionary history and improve our ability to identify the genetic causes of human disease. With the advent of exome and genome data sets of parent-offspring trios we have an unprecedented opportunity to characterize mutations at an individual level and to harness the increasing sample sizes to identify disease-causing mutations.

1.2 Mutational processes

1.2.1 Types of mutation

Mutations can cause small or large changes to DNA. When a mutation causes a single base pair change this leads to a single nucleotide variant (SNV). Occasionally a single mutational event can create multiple base pair changes in close proximity (typically within 20bp). This is referred to as a multi-nucleotide variant (MNV). When SNVs and MNVs fall within protein coding sequences they can lead to changes in the amino acid sequence which can alter the protein product (missense variants) or lead to a premature truncation of the protein (protein truncating variants (PTVs)).

Mutations can also result in insertions or deletions (indels). Indels range widely in size. They can affect just one or two base pairs up to tens of thousands of base pairs which can affect entire genes. When small indels fall into coding sequences they can shift the reading frame, leading to a frameshift or if they are divisible by three will lead to an inframe indel.

Structural variants encompass larger types of genetic variation including large insertions, deletions, inversions or duplications .

1.2.2 Origins of mutation

Mutations arise primarily from errors in DNA replication and from chemical damage to DNA. During replication, misincorporated nucleotides occasionally escape detection by proofreading mechanisms and can lead to single base changes. These single base mutations can be classified as either transitions or transversion and the rates of transitions is twice that of tranversions [129]. In addition to single base pair changes, small indels can be created by slippage of the polymerase during replication, typically in repeat regions.

Chemical damage to DNA can be induced by both endogenous and exogenous sources. DNA is vulnerable from alkylation and oxidations and can also incur spontaneous damage from hydrolysis and deamination. A common endogenous source of mutation is a result of hydrolytic deamination of cytosine which spontaneously deaminates to uracil. Uracil is easily identified as an unnatural base and can be efficiently repaired by the base excision pathway. However 5-methylcytosine can undergo deamination to Thymine which is repaired by less efficient pathways[37]. In humans, the $5'$ C in a CpG context is usually methylated and has a mutation rate that is higher than any other context[45]. The most commonly occurring endogenous nucleotide base lesion has been shown to be due misincorporation of ribonucleotides which can lead to genome instability and subsequently mutation if insufficiently repaired[176, 108].

Exogenous sources of damage to DNA can be due to various environmental mutagens. There are several well characterised examples. For example, exposure to UV radiation can create DNA lesions which then results in mutation. Ionising radiation is a well known mutagen that causes double strand breaks in DNA. Benzo(a)pyrene is a known carcinogen found in tobacco smoke which induces mutations after forming covalent DNA adducts.

1.2.3 DNA damage tolerance and repair

Repair and tolerance mechanisms exist to counteract DNA damage and to correct or mitigate the impact of damage. There are five primary DNA repair pathways: base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR), homologous recombination (HR) and non-homologous end joining (NHEJ). These pathways protect against specific types of damage. For example BER and NER correct damage that has occurred to single bases. To remove this damage BER typically excises a single base while NER removes a patch of nucleotides. For specific subsets of base damage, such as UV photolesions and alkylated bases, there are also enzymes that can directly reverse the associated DNA damage. MMR targets base mismatches and insertion/deletion mismatches that arise as a result of DNA replication errors. HR and NHEJ are both focused on repair of double strand breaks (DSBs). Homologous recombination repairs DSBs by utilising DNA sequence information from homologous sequences while NHEJ repairs DSB by directly joining the broken ends. Translesion synthesis (TLS) is an example of a DNA damage tolerance mechanism. TLS is conducted by specific polymerases that allow DNA replication to proceed past aberrant DNA lesions.

These repair processes are crucial in preserving genetic stability however in certain instances can also create their own mutations. For example repair of DSBs by NHEJ typically introduces errors, often indels, at the repair site. Additionally translesion synthesis polymerases (such as REV1, POLζ, POLη,POLκ and POL ι) bypass DNA lesions in order to continue replication but frequently introduce an incorrect base in the process [96, 132, 216, 160].

Defects in DNA repair pathways can lead to an increase in mutation rate, which in somatic tissue can lead to cancer. Many known cancer driver mutations fall within genes involved in these repair processes [217, 28]. Germline variants in certain DNA repair genes can predispose individuals to cancer. For example germline variants in MMR genes are known to cause Lynch syndrome, a form of hereditary colorectal cancer, and variants in BRCA1/2, which are involved in homologous recombination, are known to predispose carriers to breast cancer[167]. While the link between defects in DNA repair and somatic hypermutation is well established, the effect in germline tissue is still not clear.

Different mutational mechanisms can leave distinct mutational patterns. These combinations of mutation types are referred to as ’mutational signatures’ [159]. For example, UV light is a source of DNA damage that can crosslink adjacent pyrimidine bases. If this pyrimidine dimer is not repaired, DNA polymerases will typically insert two adenines opposite the dimer resulting in a predominantly C to T or CC to TT mutational signature. More complex mutational signatures can be extracted computationally and many of these have been well characterised. There are currently >100 somatic mutational signatures that have been identified in a wide variety of cancers [4]. In cancer these can provide evidence of what gene may be defective; approximately half of these signatures have been attributed to endogenous mutagenic processes or specific mutagens [165, 5, 4].

1.3 Estimating human germline mutation rates

1.3.1 Early strategies to detect mutations

Before the ability to observe *de novo* mutations (DNMs) directly, researchers developed a range of strategies to detect mutations indirectly. The initial strategy depended on phenotypic markers that were easily observed.

The first estimates of germline mutation rates were not made in humans but in *Drosophila* and maize [154, 207, 155]. A major challenge in making these estimates came from the fact that mutations happen at a very low rate. In these organisms, estimates were made from breeding many individuals and then scoring large numbers of offspring. In *Drosophila* this was done by counting the number of lethal mutations, which would impact the number of viable offspring, as an approximation of the mutation rate. In maize the mutation rate was estimated using several easily observed traits. This approach is, of course, not feasible in humans and so another strategy was needed. Haldane made one of the first estimates of the human mutation rate in 1935 by using estimates of the frequency of haemophiliac men in London [72]. This approach relied on the idea that the frequency of the harmful allele was a function of the balance between the mutation rate and the resulting fitness of that allele. Subsequent estimates were calculated using other monogenic diseases [109]. These methods made several assumptions which may affect their accuracy. Firstly they had to make an estimate of the mutational target for the disease and assumed complete ascertainment of the phenotype. They were also only able to base their mutation rate estimate on coding regions of the genome.

The mouse specific-locus test was developed in the 1950s [23, 184]. This initially involved exposing a wild-type animal to a mutagen, such as ionising radiation, and crossing them with homozygotes for specific recessive genes that were easily visible (such as ear type, hair colour or certain eye traits). The first generation offspring are then examined for differences from the expected wild type as evidence of mutation[185]. This required a large number of mice however the method was relatively straightforward with regards to counting mice with a visible phenotype. This test was able to detect specific recessive mutations and visible dominant mutations. Detecting dominant mutations depended on the researchers' power of observation and was rather subjective. Mutations affecting certain physical traits could be obvious but mutations affecting behaviour or small physical changes could easily be missed[192]. The background and radiation-induced mutation rates for 8 specific loci were estimated in the 1950s and required ~85,000 mice in total. This allowed for a better estimate of the impact of radiation on humans than previous experiments conducted in *Drosophila*.

Another strategy for estimating mutation rates came not from direct counting of phenotypes but through evolutionary comparisons. The basis of this approach is that the mutation rate for neutral mutations is equal to the rate of evolution [106, 155]. This means that one can calculate the amount of sequence divergence between non-coding DNA sequences of two species and if one has a good estimate of the time at which these species diverged then one can use these two pieces of information to estimate the mutation rate. This strategy was applied by comparing pseudogenes between human and chimpanzees and the mutation rate was estimated to be ~2.5×10^{-8} mutations per base per generation[156, 110].

1.3.2 Methods for recent direct mutation rate estimates

Advances in available technology allowed mutation rates of specific types of genomic variation to be estimated directly. With the advent of DNA fingerprinting in 1985, the mutation rate of tandem-repetitive ’minisatellite’ regions in the human genome was one of the first examples of a human mutation rate estimated directly from a pedigree [95, 93]. This was done from a dataset of 40 large families that consisted of 344 offspring which allowed for identification of *de novo* changes in the length of minisatellites. Locus specific mutation rates were estimated from PCR assays of pooled sperm DNA [94]. Two decades later, CNVs were found to be widespread in the human genome and constituted an important source of genomic variation.[194, 87]. Locus specific mutation rates were, like minisattelites, initally estimated from pooled sperm DNA [218, 227]. The development of the CNV microarray allowed for identification of *de novo* CNVs from parent-offspring trios which yielded initial direct estimates of the CNV mutation rate [193, 239, 90].

The emergence of massively parallel sequencing technologies has now allowed us to directly observe *de novo* SNVs and indel mutations in pedigrees. Recent whole-genome and exome sequencing studies have allowed us to sequence trios as well as larger family structures. These studies have estimated the mutation rate to be ~1.2×10^{-8} per base per generation for SNVs. This corresponds to ~70 DNMs per individual [33, 180, 111]. This estimate is almost half that of the mutation rate estimates derived from earlier approaches. Discrepancies between these estimates affect our understanding of the timing of human evolution and may suggest that the mutation rate has changed over time [190]. More generally, it highlights a need to understand potential sources of variability and/or error in mutation rate estimates.

1.4 Variation in the human germline mutation rate

Variation in the human germline mutation rate can be considered from different perspectives: within genomes, between individuals and between populations.

1.4.1 Variation within genomes

The rate of mutation for SNVs varies by several orders of magnitude along the genome and several factors appear to affect this (Figure 1.1). Sequence context is an important source of variability. The bases flanking either side of a nucleotide, the tri-nucleotide context, have varying mutation rates beyond the increased mutability in CpG sites [86, 16]. More recent work has also shown that this variability extends beyond a trinucleotide context and there is additional value in considering the heptanucleotide context [3, 22]. On a slightly broader scale, the surrounding context of CpGs impacts their mutation rate. GC content around the CpG appears to increase the stability of methylated cytosine. This, in turn, appears to reduce its mutation rate[80]. DNA replication timing is another source of variability. Studies have shown a higher mutation rate in late-replicating regions in the germline, perhaps due to a depletion of free nucleotides[208]. Mutations are also more likely to occur near recombination hot-spots[127]. A recent study estimated that the mutation rate is ~50 fold higher within 1 kb of crossovers. They also observed that females, but not males, have increased mutation rates up to 40 kb from crossovers, especially if these are complex [73]. Transcription also affects mutational patterns. In transcribed regions, the mutation rate of A>G and A>T substitutions is higher on the non-coding strand compared to the coding strand which is most likely due to transcription coupled repair [52, 153, 68]. Low complexity regions (LCRs) have been found increase the mutation rate in the surrounding DNA. This increase in mutation rate has been found to correlate with the distance from the LCR [126].

Chromatin and nucleosome organisation may also affect mutation rates across the genome however their role is not well established. *De novo* SNVs are more abundant in regions of closed chromatin however this could be due to the fact that open chromatin is more accessible for a range of different DNA repair activities and is associated with high transcription rates and allows transcription-coupled repair to act in these regions [138]. The role of nucleosome organisation on mutation rate is also unclear. For example high nucleosome occupancy was initially thought to reduce nearby mutation rate [146] however a more recent study found the opposite effect in a different dataset [204]. Recent work has also found elevated mutation rates around translationally stable nucleosomes[128].

Another source of genomic variation in mutation rates is dependent on their relative position to each other. Mutations may not always occur independently. Mutations appear to

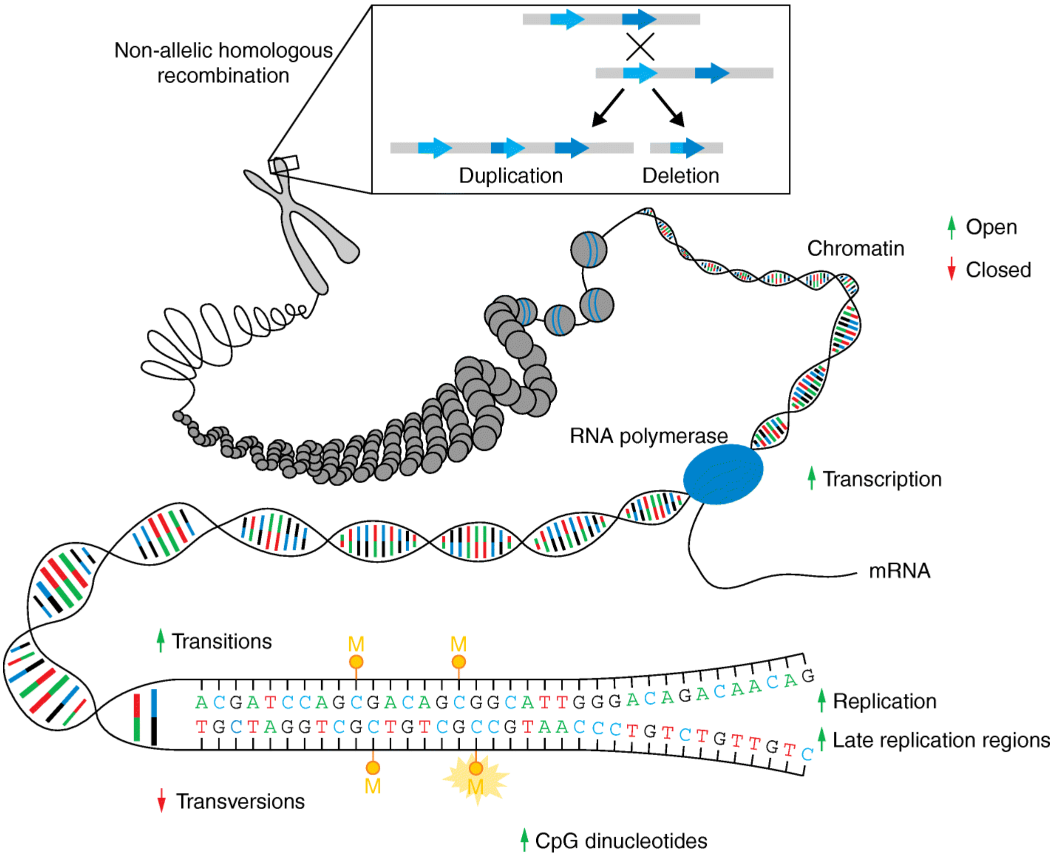


Fig. 1.1 Mechanisms of and genome level factors that influence the rate of *de novo* mutations. Figure sourced from Acuna-Hidalgo *et al* [1]

be more clustered in the genome than we would expect if they were independent [146, 196]. This has been observed for distances of up to 20 kb. These mutations do not appear to have different properties with respect to recombination rate or replication timing [52]. On a finer scale, there is an excess of pairs of mutations within 100bp that appear to be in perfect linkage disequilibrium in population samples [195, 209, 76]. A subset of these clustered mutations which fall within 20 bp of each other are referred to as multi-nucleotide variants (MNVs). The mechanisms driving MNVs are explored in more depth in Chapter 2.

Specific DNMs have been found to be highly recurrent. Moreover, the prevalence of these mutations strongly correlates with increasing paternal age. These DNMs appear to confer a selective advantage within spermatogonial stem cells which lead to clonal expansion within the testis [64, 66]. Well-known examples of these are mutations include those that fall in genes such as *FGFR2*, *FGFR3*, *RET*, *PTPN11* and *HRAS*. These mutations also lead to congenital skeletal disorders and can increase cancer risk[66, 64]. Since these mutations are positively selected for in the testis but are deleterious to the organism they have been termed

’selfish mutations’ [64, 1]. The disorders caused by these ’selfish’ mutations have an incidence up to 1000 times greater than expected based on the mutational target size[67, 10, 1, 65].

Aside from SNVs, variation in mutation rate of other forms of genetic variation has also been observed. Mutational hotspots have been identified for copy number variants (CNVs) which have up to a ~100 fold increase of mutation rate. These hotspots are enriched for being located near to segmental duplications likely due to these CNVs arising from non-allelic homologous recombination[175, 53]. Variation across the genome has also been observed for short tandem repeats (STRs)[49, 70]. Studies have observed a positive association between repeat number and mutation rate in humans[84]. Length of the repeat unit has also been shown to influence the STR mutation rate[210, 70]. For indels, it has been shown that reptitive regions, including STRs and homopolymer runs, increases the indel mutation rate due to the higher propensity of polymerase slippage in these regions [55, 151, 121]. The increase in the indel mutation rate in repetitive regions has shown to depend on both the size of the repeat unit and the length of the repeat tract[151].

1.4.2 Individual level variation

Mutation rates vary between individuals and several factors that influence this rate have been identified. Approximately 3-4 times as many mutations originate from the father than the mother which indicates differences in male and female germline mutation rates[71, 173, 111]. The number of DNMs observed in an individual is highly associated with paternal age. Paternal age accounts for the majority of mutation rate variation between individuals[111, 98]. It has been estimated that there is an increase of ~2 DNMs for every additional year in father’s age [111, 173]. It has been proposed that the increasing number of cell divisions with age in the male germline is a likely source of this effect. In women, oocytes undergo a fixed number of cell divisions early in their life. In men, spermatogonid stem cells replicate continuously throughout their life allowing for more replicative errors (Figure1.2) . A recent study interrogating this hypothesis has examined the fraction of paternal mutations in phased DNMs and has found that this does not increase with paternal age[54]. This may suggest that replication is not driving the paternal age effect and the mutations are predominantly damage-induced. Interestingly, during the course of my PhD, a maternal age effect has also been detected with a more subtle increase of ~0.5 DNMs/year [234, 98]. This could be due to an accumulation of spontaneous mutations over time in the female germ cells. This parental age effect has also been shown to differ between families. A study of three families, each with 4 or 5 children, has shown that the paternal age effect may differ across families [173]. A more recent analysis of 33 families of three generations from Utah has confirmed significant differences between parental age effects [189].

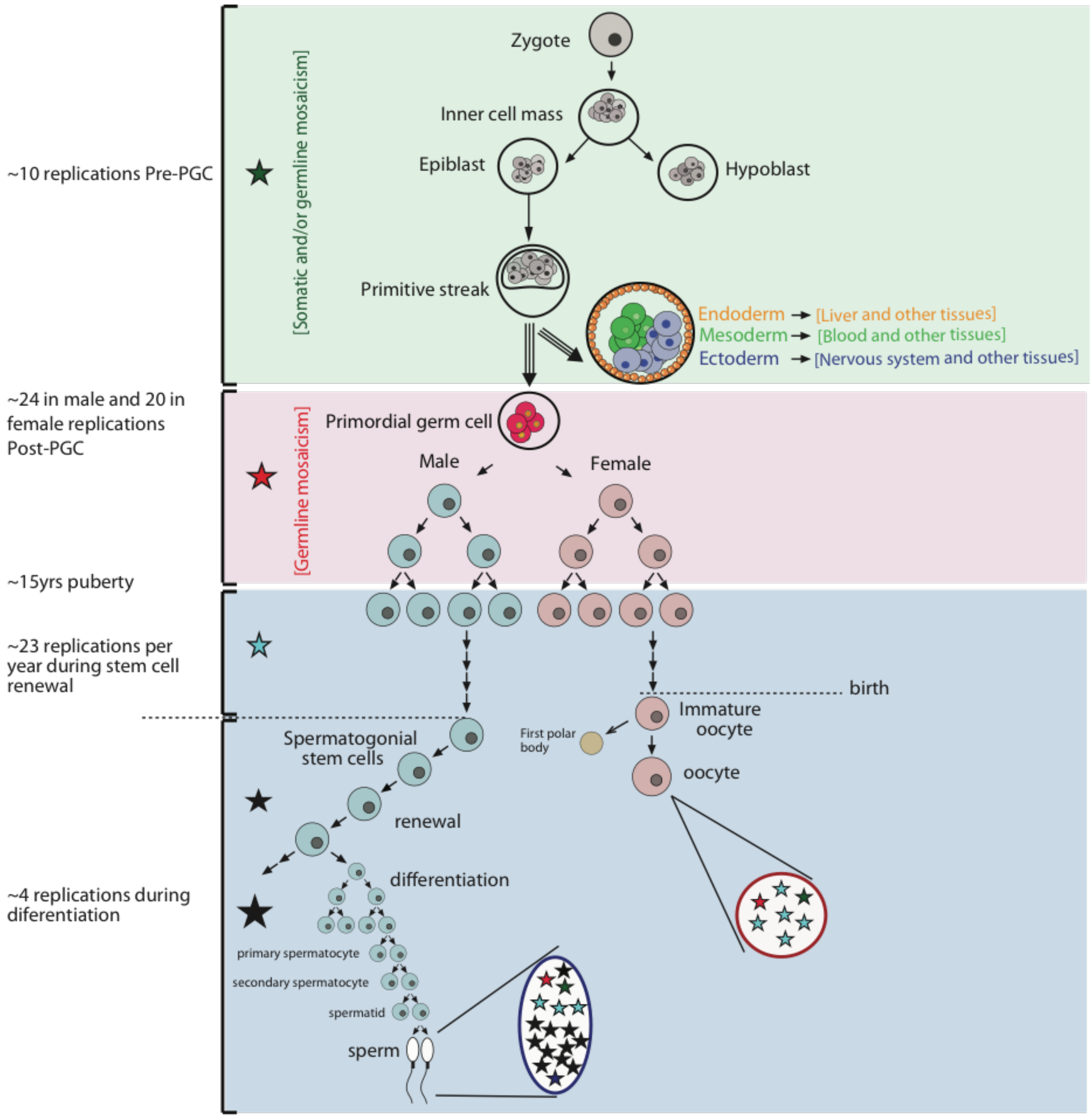

~10 replications Pre-PGC
[Somatic and/or germline mosaicism]
Zygote
Inner cell mass
Epiblast
Hypoblast
Primitive streak
Endoderm → [Liver and other tissues]
Mesoderm → [Blood and other tissues]
Ectoderm → [Nervous system and other tissues]
~24 in male and 20 in female replications Post-PGC
[Germline mosaicism]
Primordial germ cell
Male
Female
~15yrs puberty
~23 replications per year during stem cell renewal
birth
Immature oocyte
First polar body
oocyte
Spermatogonial stem cells
renewal
differentiation
~4 replications during diferentiation
primary spermatocyte
secondary spermatocyte
spermatid
sperm

Fig. 1.2 Embryogenesis and gametogenesis. Figure sourced from Rahbari *et al* [173]

The differences in parental age effects between mothers and fathers prompts the question of what differences there may be in the mechanisms generating these mutations. Differences in the mutational spectra between maternal and paternal mutations can help inform what these could be. Subtle, but significant, differences have been observed between the overall mutational spectra[62]. Some of these differences become more pronounced with increased parental age [98, 54, 62]. The types of *de novo* maternal mutations change with the mother's age. Specifically a significant decrease in CpG>TpG mutations and an increase in C>G mutations has been observed. It has been hypothesised that this increase in C>G maternal mutations, which does not occur in paternal mutations, is associated with double-strand breaks in aging oocytes[54, 62]. As mentioned previously, the paternal and maternal mutation rates appear to be different around crossover sites. This is likely due to the fact that the location of recombination hotspots differs between males and females and that the recombination rate differs at shared sites[15]. Beyond this, the mutational spectra of the DNMs around these sites are also significantly different [73]. It has been suggested that this may be due to differences in the sex-specific timing of meiosis in the germline development but this is still unclear. As sample sizes increase in trio-based studies a larger collection of DNMs will be crucial in examining these parental differences in more detail.

The timing of mutations is an important source of individual variation. Mutations in the germline can occur at any stage of development from zygote to gamete. When mutations occur prior to the specification of primordial germ cells (PGCs) they can result in mosaicism across somatic tissues. At least 3-4% of *de novo* mutations found in offspring have been found to be mosaic in parental somatic tissues (Figure1.2) [173, 189]. Mutations that occur around the time of PGC specification may result in germline mosaicism. These can result in an increased probability of siblings sharing the same *de novo* mutation. Mutations that occur early in the first few cell divisions in embryogenesis in the offspring can result in post-zygotic mutations (Figure1.2). The mutant allele proportion for such mutations should be lower than 50% and may be detected with incomplete sensitivity in sequencing data[99]. The contribution of mutations at each stage to the overall mutation rate is not well understood but is important in understanding mutational processes and understanding variation within families. This can be critical for estimating the risk of recurrence for families with diseases caused by damaging *de novo* mutations.

Other sources of variation in individual mutation rate could be due to genetic differences influencing mutation rate or spectra (e.g. in DNA repair pathways) or to DNA damage caused by differential exposure to mutagens. For example tobacco smoke has been shown to contribute to paternal germline mutations in mice, although a similar effect has yet to be observed in humans[242].

1.4.3 Population level variation

Mutation rates have evolved over time just as any other phenotype. This is most clearly demonstrated by the variation in mutation rates between species [152, 135]. The mutation rate in mice has been observed to be higher than in humans and differences have been seen in mutation spectra and mutation rates per cell division[35, 30, 130]. The mutation rate per spermatogonial stem cell (SSC) division is estimated to be lower in humans in mice, this is hypothesised to have evolved as a result of the larger contribution of SSC to the human germline[130]. Since the split of hominoids and monkeys the per year mutation rates have decreased in hominoids, known as the hominoids slow down. It has been observed that evolutionary rates are faster in new world monkeys compared with old world monkey and in turn rates in old world monkeys are faster than in humans and apes[152, 63, 206]. On a shorter evolutionary time scale, mutation rates have also been suggested to differ between human populations. The mutational signature TCC->TTC is enriched in rare variation within European populations[75]. There have also been signatures shown to be private to certain Native American populations [142]. The reasons behind these differences is still elusive. It has been suggested that heterozygosity affects mutation rate and that this has impacted differences in mutational spectra across populations [240, 9]. Characterising population specific mutation rates in more detail will help to elucidate different selective pressures on the mutation rate and possible differences in underlying mechanisms.

1.5 *De novo* mutations in human disease

1.5.1 Modes of inheritance

De novo mutations are a significant cause of rare genetic disease. To understand their contribution it is important to first contextualise the possible modes of inheritance for genetic disorders in general. Diseases are referred to as ’Mendelian’ when disease-causing alleles segregate according to Mendel’s laws of inheritance. These disorders tend to be monogenic and are caused by rare and highly penetrant mutations. Autosomal recessive disorders only occur when an individual has two mutant alleles in the disease-associated gene. This usually occurs when an individual inherits a mutant allele from each parent therefore these disorders are unlikely to be caused by *de novo* mutations. Autosomal dominant disorders occur in individuals with only a single mutant allele in a disease-associated gene. This mutant allele can be inherited from a parent, who is likely affected, or it can arise *de novo* in the individual. X-linked disorders have slightly different inheritance patterns compared to those on the autosome. X-linked recessive disorders can occur in females when they inherit one disease allele from their father and one from their mother. For males with an X-linked recessive disorder, since they only have a single copy of the X chromosome, they always inherit this mutant allele from their mother. This means that males are much more likely than females to be affected by X-linked disorders as they only need a single mutant allele. X-linked dominant disorders are much rarer but there are a few examples. For example, Rett syndrome is a developmental disorder caused by dominant mutations in *MECP2*. This disorder almost exclusively affects females as it appears to be embryonic lethal in males. This disorder is almost always caused by a *de novo* mutation.

1.5.2 Historical context

Research on the contribution of genetic variation to human disease has historically been focused on inherited variation. Linkage analysis was one of the first approaches used to associate regions of the genome to human disease and was fundamental in identifying Mendelian disease genes[17]. This was done by testing if a series of marker alleles co-segregated with disease status within a family or across multiple families. The emergence of microarray technology and the completion of the reference genome, through the Human Genome Project, led to the development of genome-wide association studies (GWAS). GWAS were able to test associations between allele frequencies of hundreds of thousands of SNVs with human disease across the genome. This became crucial in the progress of complex disease genetics.

Chromosomal aneuploidies were one of the the first types of observable *de novo* genetic variation that were shown to cause a disorder. The trisomy of chromosome 21, which causes Down's syndrome, was first observed through a microscope in 1959 [124]. By the 1990s a few *de novo* SNVs and large (>1MB) CNVs were shown to cause sporadic disease, however the ability to systematically study the role of smaller scale *de novo* mutations in human disease has only been possible since the development of genomic microarrays and next-generation sequencing technologies[26, 245, 134]. Early studies using Array Comparative Genomic Hybridisation (CGH) showed that *de novo* CNVs were significantly enriched in individuals with autism, epilepsy and developmental delay [193, 40]. Moreover, specific genes were associated with sporadic disease [225, 221]. The emergence of exome and whole genome sequencing has been paramount in allowing for detection of *de novo* SNVs and indels in patients with rare genetic disease. This has led to the identification of hundreds of genes associated with rare sporadic disease [1, 18, 59]. Studies have also established that *de novo* mutations are implicated in common neurodevelopmental disorders such as autism, epilepsy and intellectual disability[89, 163, 6, 40].

1.5.3 Developmental disorders

Developmental disorders (DD) encompass neurodevelopmental disorders, congenital anomalies, abnormal growth parameters and unusual behavioural phenotypes [50]. Although these disorders are individually rare, collectively they affect 2-5% of births in the UK [198, 183, 41]. DNMs are enriched in cohorts with DD and it has been estimated that ~50% of severe DD cases are caused by a pathogenic coding DNM [140, 41]. DNMs in specific non-coding regions, such as conserved regulatory elements, are expected to explain 0.5-2% of severe DD [201]. Pathogenic DNMs in DD mostly lead to dominant genetic disorders although they can also cause recessive disorders via compound heterozygosity in conjunction with an inherited variant. There are also examples where a patient's phenotypes are explained by more than one pathogenic DNM [241, 237].

The timing of DNMs has an important impact on both the presentation and clinical impact of DDs. Post-zygotic mutations are difficult to detect but technological advances have allowed for more study into this type of mutation [20]. Post-zygotic pathogenic DNMs can result in somatic mosaicism which can lead to a less severe phenotype compared to that caused by a constitutive mutation [170, 1, 161]. There are also examples of disorders where the DNMs only appear as mosaic such as those caused by mutations in *AKT1* or *PIK3CA* [229, 113]. This may be because the mutation is lethal when constitutive [74]. Mosaic disorders also tend to be caused by activating missense mutations which result in cell proliferation and overgrowth. The clinical impact of the timing of mutation is especially important. It has been estimated that ~3% of pathogenic DNMs in DD are post-zygotic in the affected child [235]. This means that the sibling recurrence risk is very low which can be crucial information for future pregnancies. In addition, evidence of a post-zygotic mutation in a parent, which can be detected in <1% of cases, results in a much larger sibling recurrence risk [235, 97].

There are currently ~2000 genes known to be associated with developmental disorders[236]. Initially these genes were discovered using a phenotype-driven approach by aggregating patients with similar clinical presentations. However we know that the clinical manifestations of these disorders can vary widely and so in more recent years gene discovery has been complemented with a genotype-driven approach. This has involved the ascertainment of large cohorts with a diversity of related phenotypes. Novel associations between genes and DDs have then been identified by looking for genes where we observe a significantly greater number of non synonymous DNMs than expected under a null mutational model [41, 215]. Analysing DNM burden has also discovered novel gene associations for a range of different disorders aside from DDs such as autism, schizophrenia and congenital heart disease [202, 69, 157, 205].

The Deciphering Developmental Disorders (DDD) Study is an example of a large cohort that has been important in furthering our understanding of the genetic architecture of DD. The DDD study consists of 13,451 patients with a severe, but genetically undiagnosed, developmental disorder. These patients have been recruited from 24 regional genetic services across the UK and Republic of Ireland. There is a wide array of phenotypes in the dataset however 87% have intellectual disability (Figure1.3) [215]. The patients have been systematically phenotyped, half have had array CGH and all have been been exome sequenced. For approximately 88% of these patients in the DDD the parents have also been exome sequenced. This allows for interrogation of *de novo* variation. In 2017, the DDD published results on analysis of *de novo* variants in 4,293 patients and identified 94 significant genes associated with DD, 14 of which were not previously known[41]. The exome-sequencing data generated as part of the DDD study are used in all three results chapters in this thesis.

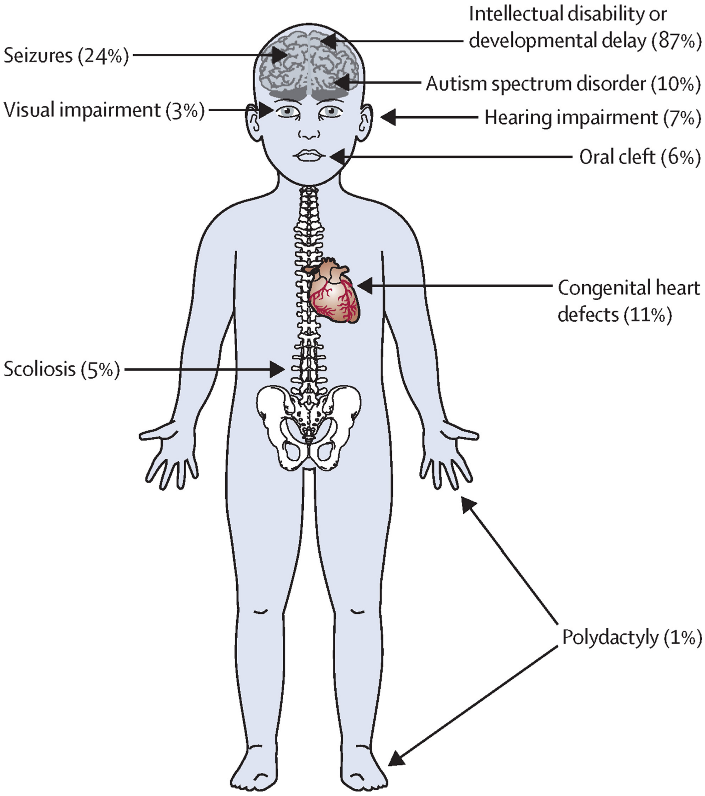

Intellectual disability or developmental delay (87%)
Seizures (24%)
Autism spectrum disorder (10%)
Visual impairment (3%)
Hearing impairment (7%)
Oral cleft (6%)
Congenital heart defects (11%)
Scoliosis (5%)
Polydactyly (1%)

Fig. 1.3 Phenotypes in the DDD study. Figure sourced from Wright *et al*[236]

The 100,000 Genomes Project is a more recent initiative in the UK to whole-genome sequence patients with rare diseases and cancer. Genomics England Ltd was a company set up by the Department of Health and Social Care to deliver the project. As of December 2018, the milestone of 100,000 genomes was reached and Genomics England have expanded their goal to sequence 5 million genomes in the next five years. The rare disease arm of the 100,000 Genomes Project currently consists of sequencing data from ~72,000 participants which reflects ~34,000 families with a variety of family structures, as described in Table 1.1 (as of April 2020). The project includes a wide array of more than 190 rare diseases, the breakdown of the types of these diseases are described in Table 1.2. Approximately 40% of these patients are described as having a neurodevelopmental disorder. Data from the project

Rare Disease Family type	Count
Duo with Mother or Father	4,797
Duo with other Biological relative	1,286
Families with more than three participants	1,856
Singleton	12,607
Trio with Mother and Father	11,854
Trio with Mother or Father and other Biological Relationship	829
Trio with other Biological Relatives	384
TOTAL	33,613

Table 1.1 Distribution of family types within the rare disease arm of the 100,000 Genomes Project

Normalised Disease Group	Count	Proportion
Cardiovascular disorders	3,799	0.111
Ciliopathies	340	0.010
Dermatological disorders	389	0.011
Dysmorphic and congenital abnormality syndromes	591	0.017
Endocrine disorders	849	0.025
Fastroenterological disorders	118	0.003
Frowth disorders	186	0.005
Haematological and immunological disorders	913	0.027
Hearing and ear disorders	801	0.023
Infectious diseases	13	0.000
Metabolic disorders	694	0.020
Neurological and neurodevelopmental disorders	14,095	0.411
Opthalmological disorders	2,866	0.084
Psychiatric disorders	83	0.002
Renal and urinary tract disorders	3,576	0.104
Respiratory disorders	322	0.009
Rheumatological disorders	258	0.008
Skeletal disorders	833	0.024
Tumour syndromes	1,682	0.049
Ultra-rare disorders	1,847	0.054
TOTAL	34,255	

Table 1.2 Distribution of disease types in the rare disease arm of the 100,000 Genomes Project

were made available to researchers in 2019. The benefit of whole-genome sequencing over exome sequencing allows for more interrogation of the non-coding regions of the genome. From a mutational perspective it can also allow us to call *de novo* mutations across the whole genome which gives us power to look more closely at variation in mutational spectra and rates across individuals. Analyses from chapter 3 of the thesis are largely conducted using this dataset.

1.6 Outline of dissertation

The goal of this thesis was to understand sources of variation in germline mutation and the contribution of these mutations to rare developmental disorders. These sources of variation encompassed types of mutations that have been previously underrepresented in genetic research as well as individual mutation rates and spectra across individuals and parental origin.

Chapter 2 of this dissertation focuses on the mutational origins and pathogenic impact of MNVs. Here I describe how MNVs in protein-coding sequences can be more pathogenic than an SNV even when the MNV falls within a single codon. I also estimate the MNV mutation rate, explore the mutational spectra of these variant and describe the contribution of *de novo* MNVs to severe developmental disorders. This analysis was conducted on trio exome data from the DDD study.

Chapter 3 of this dissertation focuses on identifying and characterising germline hypermutators. Using whole-genome sequencing data from the DDD and GEL dataset, I identified fifteen children with an unusually large number of *de novo* mutations. Eight of these appear to be due to a paternal hypermutator. I describe analyses to try and identify a genetic cause for this hypermutation of which I found two putative paternal variants in DNA repair genes. I will also describe work focussed on whether variants in DNA repair genes impact germline mutation rates by examining a well characterised cancer somatic mutator gene and then using a broader approach across all DNA repair genes. Using the large resource of DNMs called in the GEL dataset, I also explore other sources of variation in germline mutation rate, including differences between maternal and paternal DNMs as well as the effect of parental age.

In **Chapter 4**, I describe analyses that focussed on using *de novo* mutations to identify novel genes associated with DD. This was performed with exome parent-offspring data from the DDD study pooled with trios from GeneDx, a US-based genetic diagnostic company, and trios from Radboud University Medical Center (RUMC). This chapter also describes work done to explore how these novel genes differ from those that were previously known, as well as a model-based approach to explore the likely properties of currently undiscovered genes.

Lastly, in **Chapter 5**, I summarise the main findings of these projects and what can be learnt from them. This then leads on to a discussion of how this work can be extended and what developments can advance our knowledge of human germline mutation in upcoming years.

Chapter 2

Exome-wide assessment of the functional impact and pathogenicity of multinucleotide mutations

2.1 Introduction

In genomic analyses, single nucleotide variants (SNVs) are often considered independent mutational events. However SNVs are more clustered in the genome than expected if they were independent [146, 196, 8]. On a finer scale, there is an excess of pairs of mutations within 100 bp that appear to be in perfect linkage disequilibrium in population samples [195, 209, 76]. While some of this can be explained by the presence of mutational hotspots, natural selection or compensatory variants, it has been shown that multi-nucleotide mutations play an important role [191]. Recent studies found that 2.4% of *de novo* SNVs were within 5 kb of another *de novo* SNV within the same individual [14], and that 1.9% of *de novo* SNVs appear within 20 bp of another *de novo* SNV [191]. Multi-nucleotide variants (MNVs) occurring at neighbouring nucleotides are the most frequent of all MNVs [14]. Moreover, analysis of phased human haplotypes from population sequencing data also showed that nearby SNVs are more likely to appear on the same haplotype than on different haplotypes [191].

The mutational origins of MNVs are not as well understood as for SNVs, however different mutational processes leave behind different patterns of DNA change which are dubbed mutational ‘signatures’. Distinct mutational mechanisms have been implicated in creating MNVs. Polymerase ζ is an error-prone translesion polymerase that has been shown to be the predominant source of *de novo* MNVs in adjacent nucleotides in yeast[76, 14]. The most common mutational signatures associated with polymerase ζ in yeast have also been observed to be the most common signature among MNVs in human populations[76], and were also found to be the most prevalent in *de novo* MNVs in parent-offspring trios [14]. It has been suggested that translesion DNA polymerases play an important role in the creation of MNVs more generally [27]. A distinct mutational signature has also been described that has been attributed to the action of APOBEC deaminases [5].

Although MNVs are an important source of genomic variability, their functional impact and the selection pressures that operate on this class of variation has been largely unexplored. In part, this is due to many commonly used workflows for variant calling and annotation of likely functional consequence annotating MNVs as separate SNVs [188]. When the two variants comprising an MNV occur within the same codon – as occurs frequently given the propensity for MNVs at neighbouring nucleotides – interpreting MNVs as separate SNVs can lead to an erroneous prediction of the impact on the encoded protein. The Exome Aggregation Consortium (ExAC) systematically identified and annotated over 5,000 MNVs that occurred within the same codon in genes, including some within known disease-associated genes[125]. Although individual pathogenic MNVs have been described [115], the pathogenic impact of MNVs as a class of variation is not yet well understood.

2.1.1 Chapter overview

In this chapter I analysed 6,688 exome sequenced parent-offspring trios from the Deciphering Developmental Disorders (DDD) Study to evaluate systematically the strength of purifying selection acting on MNVs in the population sample of unaffected parents, and to quantify the contribution of pathogenic *de novo* MNVs to developmental disorders in the children.

2.1.2 Publication and contributions

The results described in this chapter were published in 2019 [100]. I briefly summarise the various contributions to this project. Giuseppe Gallone performed the upstream variant calling for the DDD project, Jeremy McRae called and filtered the *de novo* mutations (DNMs) in DDD and Elena Prigmore experimentally validated the *de novo* MNVs. All of this work was done under the supervision of Matthew E. Hurles. The parts of the publication reproduced in this Chapter are all my original work.

2.2 Methods

2.2.1 Variant and *De Novo* calling in DDD

The analysis in this chapter was conducted using exome sequencing data from the DDD study of families with a child with a severe, undiagnosed developmental disorder. The recruitment of these families with developmental disorders has been described previously[236]. 7,833 parent-offspring trios from 7,448 families and 1,791 singleton patients (without parental samples) were recruited at 24 clinical genetics centres within the United Kingdom National Health Service and the Republic of Ireland. Families gave informed consent to participate, and the study was approved by the UK Research Ethics Committee (10/H0305/83, granted by the Cambridge South Research Ethics Committee and GEN/284/12, granted by the Republic of Ireland Research Ethics Committee). In this analysis, I only included trios from children with apparently unaffected parents in our analysis to avoid bias from pathogenic inherited MNVs. This was defined as those trios where the clinicians did not report any phenotypes for either parent. This resulted in a total of 6,688 complete trios. Sequence alignment and variant calling of single nucleotide variant and insertions/deletions were conducted as previously described. DNMs were called using DeNovoGear and filtered as described in a previous DDD publication [41, 174].

2.2.2 Estimating the MNV mutation rate

I estimated the MNV mutation rate by scaling the SNV mutation rate estimate of 1.1×10^{-8} mutations per base pair per generation by the ratio of MNV segregating sites/ SNV segregating sites observed in our data set[180]. This approach is based on a rearrangement of the equation for the Watterson estimator[226]. This is outlined below where θ is the Watterson estimator, μ is the mutation rate, K denotes the number of segregating sites, N_e is the effective population size, n is the sample size and a_n is the $n-1$th harmonic number.

$$\begin{aligned}
\hat{\theta} &= \frac{K_{SNV}}{a_n} = 4N_e\mu_{SNV} \\
\mu_{SNV} &= \frac{K_{SNV}}{a_n 4N_e} = 1.1 \times 10^{-8} \\
a_n 4N_e &= \frac{K_{SNV}}{1.1 \times 10^{-8}} \\
\mu_{MNV} &= \frac{K_{MNV}}{a_n 4N_e} \\
&= \frac{K_{MNV}}{K_{SNV}} 1.1 \times 10^{-8}
\end{aligned}$$

To avoid any potential bias from selection I excluded variants that fell into potentially constrained genes (pLI>0.1).

To ensure the validity of this method, I also estimated the SNV missense mutation rate in the same way by scaling the overall SNV mutation rate by the ratio of the number of missense SNVs in unconstrained genes compared to all SNVs and obtained an estimate of the missense mutation rate across coding regions to be 1.07×10^{-8} per coding base pair per generation which agrees with the estimate of 1.09×10^{-8} per coding base per generation which was calculated using the trinucleotide context mutational model as described by Samocha et al[187].

2.2.3 Estimating the enrichment of *de novo* MNVs

To test for the enrichment of *de novo* MNVs I used a Poisson test for three categories of genes: all genes, genes known to be associated with developmental disorders and genes that are not known to be associated with developmental disorders. Genes known to be associated with developmental disorders, in which *de novo* mutations can be pathogenic, were defined as those curated on the Gene2Phenotype website (htttp://www.ebi.ac.uk/gene2phenotype/) and listed as monoallelic that were ‘confirmed’ and ‘probable’ associated with DD. I did the same tests for synonymous, missense and protein-truncating variants using gene-specific mutations rates for each consequence type derived by Samocha et al, 2014 [187]. Significance of these statistical tests was evaluated using a Bonferroni corrected p-value threshold of 0.05/12 to take into account the 12 tests across all three subsets of genes, SNV consequence types and MNVs. To correct for sequence context when comparing DD genes and non-DD genes, I adjusted the expected number of MNVs in the DD genes category based on the excess of polymerase ζ dinucleotide contexts.

2.2.4 Estimating the number of clinically reported MNVs

I downloaded all clinically reported variants from the website ClinVar and subsetted these variants to those that fell into autosomal dominant DDG2P genes and those that were annotated as ‘definitely pathogenic’ or ‘likely pathogenic’. This set was then subsetted to 321 genes with at least one pathogenic missense mutation. This was to ensure that missense mutations cause disease in these genes. I then counted the numbers of SNV missense variants and used this to estimate the number of expected missense MNVs across those genes. This was scaled using the ratio of the SNV to MNV missense mutation rate across these genes. The MNV mutation rate used for this calculation was specifically for $\mathrm{MNV}_{\mathrm{1bp}}$ ($\mu_{MNV_{1bp}} = 8.76 \times 10^{-11}$ mutations per base pair per generation). The $\mathrm{MNV}_{\mathrm{1bp}}$ missense mutation rate was calculated as:

$$\mu_{\text{DDG2P MNV missense}} = \mu_{MNV_{1bp}} \times \frac{2}{3} \times 0.97 \times \sum \text{coding bp in DDG2P genes}$$

Where 2/3 is the probability of a coding MNV falling within a codon and 0.97 is the probability that a within-codon MNV results in a missense change. The probability of an MNV falling within a codon was calculated from the properties of codons and the probability of a within codon MNV resulting in a missense change was calculated by looking at the consequences of all possible within-codon MNVs and calculating the proportion that result in a missense. The expected number of missense MNVs in DDG2P genes was then calculated as follows:

$$\text{E(reported pathogenic missense MNVs)} = n_{\text{reported missense SNVS}} \frac{\mu_{\text{DDG2P MNV missense}}}{\mu_{\text{DDG2P SNV missense}}}$$

This assumes that the enrichment of MNV and SNV missense mutations in these genes are comparable, as I have observed in the DDD study. This yielded an expected number of 25.67 reported pathogenic MNVs compared to 22 observed reported pathogenic MNVs. To test if this difference was significant I performed a Poisson test (p = 0.55).

2.3 Results

2.3.1 Identifying and categorising MNVs

I defined MNVs as comprising two variants within 20 bp of each other that phased to the same haplotype across >99% of all individuals in the dataset in which they appear (Figure 2.1a). This definition encompasses both MNVs due to a single mutational event and MNVs in which one SNV occurs after the other. To identify all possible candidate MNVs I searched for two heterozygous variants that were within 100bp of each other in the same individual across 6,688 DDD proband VCFs and had a read depth of at least 20 for each variant. The variants were phased using trio-based phasing, which meant that the ability to phase the variants was not dependent on the distance between them. I was able to determine phase for approximately 2/3 of all possible MNVs across all individuals. Those that could not be phased were discarded. The condition of phasing impacts the allele frequency spectrum of the MNVs I could identify since the probability of being able to phase a rare MNV in at least one individual is smaller than a more common MNV. There was a significantly smaller allele frequency in the variants that could not be phased compared to those that can be phased ($p = 1.8 \times 10^{-46}$,Wilcox test). I do not expect the mutational mechanisms to differ at different allele frequencies although the rare MNVs may be more likely to be more damaging; this would make the assessment of pathogenicity more conservative. Phasing also provided an additional layer of quality assurance by requiring that the variant was called in both parent and child. MNVs tend to have lower mapping quality scores than SNVs and so traditional variant filtering criteria based on quality metrics could potentially miss a substantial number of MNVs. This also enabled me to use the same filtering criteria for different classes of variants to ensure comparability. The threshold distance of 20 bp between variants was selected as I observed that pairs of SNVs that define potential MNVs are only enriched for phasing to the same haplotype within this distance (Figure 2.1b).

De novo MNVs were defined as two *de novo* SNVs within 20bp of each other that were confirmed to be on the same haplotype using read based phasing. To identify *de novo* MNVs I looked within a set of 51,942 putative DNMs for pairs of *de novo* variants within 20 bp of each other. This set of DNMs had been filtered requiring a low minor allele frequency (MAF), low strand bias and low number parental alt reads. I did not impose stricter filters at this stage as true *de novo* MNVs tend to have worse quality metrics than true *de novo* SNVs. I found 301 pairs, approximately 1.2% of all candidate DNMs. A third of these were 1-2 bp apart (Figure 2.2a). For analysis of mutational spectra I did not filter these further however when looking at functional consequences of these *de novo* MNVs I wanted to be more stringent and examined IGV plots for all *de novo* MNVs of which 91 passed IGV examination. Ten of the *de novo* MNVs fell within genes previously associated with dominant developmental disorders. These were all validated experimentally using MiSeq or capillary sequencing. The experimental validation was done by Elena Prigmore.

In total, I identified 69,940 MNVs transmitted from the 13,376 unaffected trio parents as well as 91 *de novo* MNVs in the trio children. A set of 693,837 coding SNVs was obtained from the DDD probands with the exact same ascertainment as those for MNVs (read depth >20, phased to confirm inheritance). These were used when comparing MNV properties to SNVs to reduce any ascertainment bias.

Different mutational mechanisms are likely to create MNVs at different distances. To capture these differences, I stratified analyses of mutational spectra based on distance between the variants. The distance between the two variants that make up an MNV will be denoted as a subscript. For example, adjacent MNVs will be referred to as $\text{MNV}_{1\text{bp}}$. MNVs can be created by either a single mutational event or by consecutive mutational events. For MNVs that were created by a single mutational event, the pair of variants are likely to have identical allele frequencies as they are unlikely to occur in the population separately (I assume recurrent mutations and reversions are rare). The proportion of nearby pairs of SNVs with identical allele frequencies that phase to the same haplotype remains close to

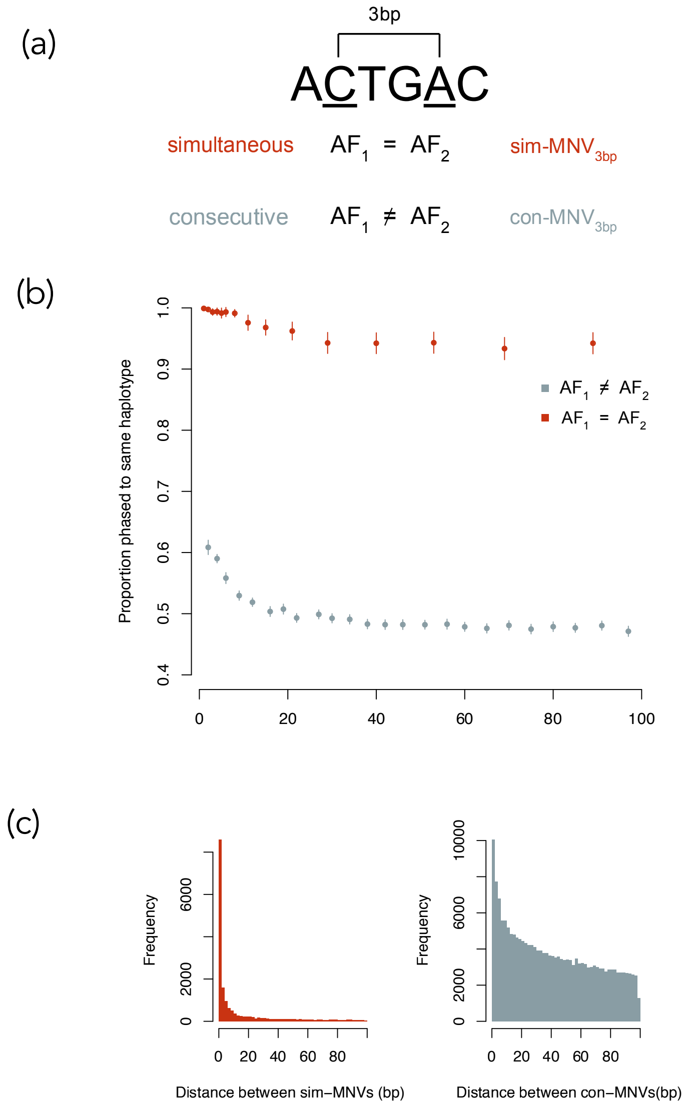


Fig. 2.1 Properties of MNVs (a) Schematic showing how sim-MNVs, two variants that occur simultaneously, are defined as having two variants with identical allele frequencies and con-MNVs, two variants that occur consecutively, as having different allele frequencies (b) Proportion of pairs of heterozygous variants (possible MNVs) that phase to the same haplotype as a function of distance separated by sim and con. (c) The number of sim-MNVs and con-MNVs by distance between the two variants.

100% even at a distance of 100 bp apart (Figure 2.1b). These variants most likely arose simultaneously and will be referred to as sim-MNVs. The proportion of pairs of SNVs with different allele frequencies that phase to the same haplotype approaches 50% at around 20bp. These SNVs probably arose consecutively and will be referred to as con-MNVs. I observed

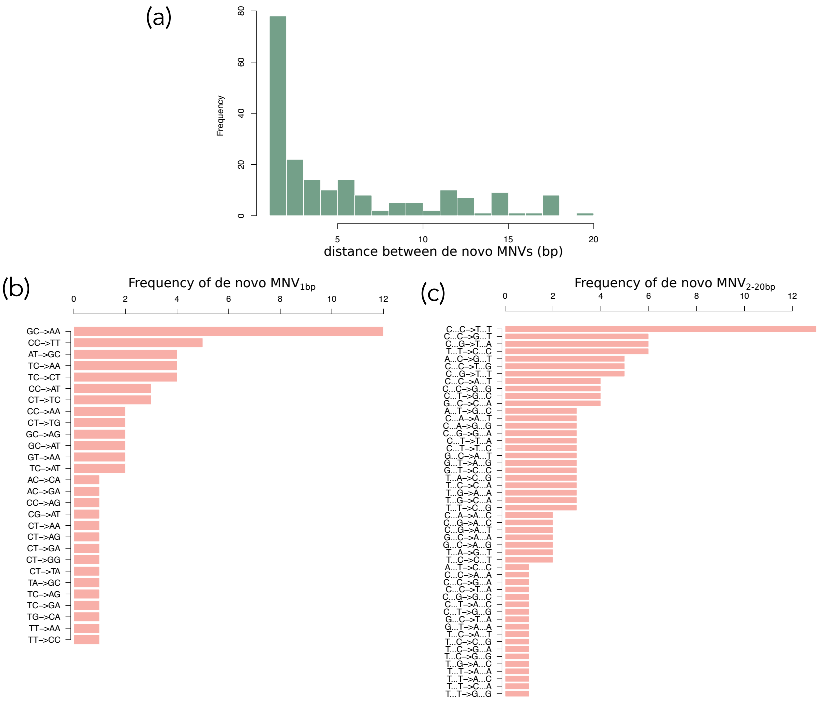


Fig. 2.2 Mutational spectra of *de novo* MNVs (a) Frequency of *de novo* MNVs according to the distance between the two variants in base pairs (b) Frequency of different mutation types for *de novo* MNV$_{1\text{bp}}$ (c) Frequency of different mutation types for *de novo* MNV$_{2\text{-}20\text{bp}}$

that sim-MNVs account for 19% of all MNVs and 53% of MNV$_{1\text{bp}}$. All *de novo* MNVs are, by definition, sim-MNVs as they occurred in the same generation.

I identified 888 trinucleotide variants (trinucleotide sim-MNVs) which I defined as three SNVs within 20 bp with identical allele frequencies. One hundred and fourteen of these occurred in three adjacent nucleotides. I observed one *de novo* trinucleotide MNV.

2.3.2 Analysis of MNV mutational spectra

Differences in mutational spectra across different subsets of MNVs can reveal patterns or signatures generated by the underlying mutational mechanism. I analysed the spectra of both simultaneous and consecutive MNV$_{1\text{bp}}$, MNV$_{2\text{bp}}$ and MNV$_{3\text{-}20\text{bp}}$. For sim-MNVs the proportion of variants that fell into these groups were 51%, 12% and 37% respectively. For con-MNVs, most variants were further away with the proportions being 10%,7% and 83% (Table 2.1, Figure 2.1c). There were significant differences between the mutational spectra of sim-MNVs and con-MNVs (Figure 2.3a,c)

DNA polymerase ζ, a translesion polymerase, is a known frequent source of *de novo* MNVs and has been associated with the mutational signatures GC->AA and TC->AA

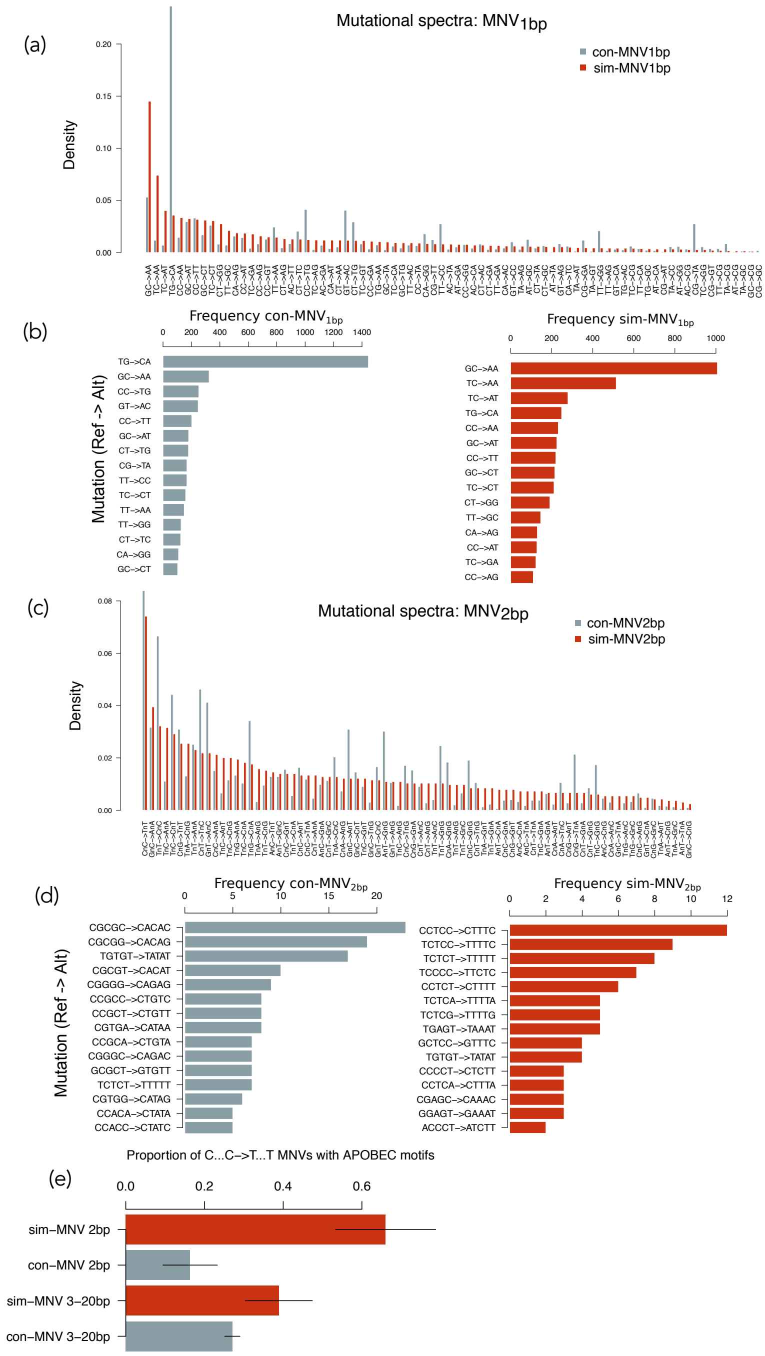


Fig. 2.3 Mutational Spectra of MNVs (a) The frequency of mutational spectra for sim-MNV$_{1bp}$ and con-MNV$_{1bp}$ (b) The 15 most common mutations for sim-MNV1bp and con-MNV$_{1bp}$ (c) The frequency of mutational spectra for sim-MNV$_{2bp}$ and con-MNV$_{2bp}$ (d) The 15 most common mutations for sim-MNV$_{2bp}$ and con-MNV$_{2bp}$ (e) The proportion of C...C->T...T MNVs that have motifs associated with mutations caused by APOBEC.

MNV type	Distance (bp)	Intra Codon	Inter Codon	Non-coding	TOTAL (% of all MNVs)
sim	1	1893	863	3850	6606 (9.4%)
	2	243	350	975	1568 (2.2%)
	3-20	-	1832	2970	4802 (6.9%)
con	1	1155	735	3923	5813 (8.3%)
	2	449	685	2649	3783 (5.4%)
	3-20	-	15316	32052	47368 (67.7%)
TOTAL (% of all MNVs)		3740 (5.3%)	19781 (28.2%)	46419 (66.4%)	69940

Table 2.1 Numbers of MNVs in each category type

[76, 14]. These signatures, and their reverse complements are the most common dinucleotide changes that I observed and account for 22% of all sim-MNV$_{1bp}$s (Figure 2.3b). These two signatures made up 18% of the *de novo* sim-MNV$_{1bp}$s which is comparable to the 20% of observed *de novo* MNVs in a recent study (Figure 2.2b) [14]. In the remaining 78% of sim-MNV$_{1bp}$s I observed sixteen other mutations, after Bonferroni multiple correction, that were significantly more prevalent in sim-MNV$_{1bp}$s compared to con-MNV$_{1bp}$s. This suggests that there are other unidentified mechanisms that are specific to creating sim-MNVs. The most prevalent sim-MNV$_{1bp}$ that is not attributed to polymerase ζ is TC>AT which accounts for 4% of all sim-MNV$_{1bp}$s. There were two *de novo* sim-MNV$_{1bp}$s with this signature however an extensive literature search, including somatic mutational signatures, has not yielded any possible mechanism behind this mutation [212].

APOBEC are a family of cytosine deaminases that are known to cause clustered mutations in exposed stretches of single-stranded DNA. These mutational signatures are commonly found in cancer and more recently discovered in germline mutations [181, 168]. The most common mutation for sim-MNV$_{2bp}$ is CnC->TnT where n is the intermediate base between the two mutated bases and is 8% of the mutations (Figure 2.3c). They are found primarily in a TCTC>TTTT or CCTC>CTTT sequence context (Figure 2.3d). CC and TC are known mutational signatures of APOBEC[77, 5, 168]. However, the APOBEC signature described previously in germline mutations was found in pairs of variants that were a larger distance apart (10-50bp). C...C -> T...T was also the most prolific mutation in sim-MNV$_{3\text{-}20bp}$ and had a significantly larger proportion of APOBEC motifs in both variants compared to con-MNV$_{3\text{-}20bp}$bp (p value 0.0056) (Figure 2.3e). The mutation C...C -> T...T was the most frequent *de novo* MNV$_{2\text{-}20bp}$ (Figure 2.2c). However only three of the twelve *de novo* MNV$_{2\text{-}20bp}$ had APOBEC motifs.

There were 6 other mutations that are significantly more common in sim-MNV_{2bp} compared to con-MNV_{2bp}. The most prevalent of these is CnG>TnT which accounts for 3% of sim-MNV_{2bp}. I did not observe any *de novo* MNVs with this mutation and I was not able to attribute a mutational mechanism after reviewing the literature.

I analysed the mutational signatures of the set of 114 adjacent trinucleotide sim-MNVs and found that the most prevalent mutation was AAA>TTT (Figure 2.4) however was not able to establish a possible mutational mechanism for this.

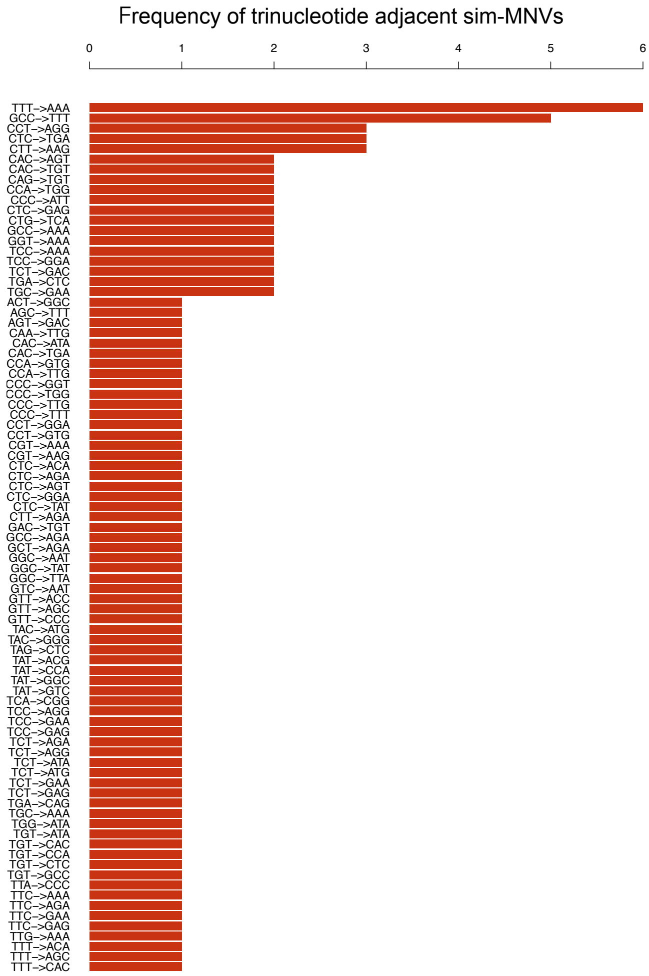


Fig. 2.4 Mutational spectra of all adjacent trinucleotide MNVs (N=114)

Mutational signatures in con-MNVs were primarily driven by hypermutability of CpG sites. In humans, the 5' C in a CpG context is usually methylated and has a mutation rate that is approximately ten-fold higher than any other context[45]. For con-$MNV_{2\text{-}20bp}$ the most common mutation is C. . . C->T. . . T and is driven by two mutated CpG sites CG. . . CG> TG. . . TG (Figure 2.3d). For con-MNV_{1bp}, 24% are accounted for by the mutation CA->TG, and its reverse complement (Figure 2.3b). These adjacent consecutive mutations most likely came about due to a creation of a CpG site by the first mutation. If the first mutation creates a CpG then the mutations would be expected to arise in a specific order: CA>CG>TG. In this scenario, the A>G mutation would likely happen first and that variant would have a higher allele frequency than the subsequent C>T. This was the case for 96% of the 1,445 CA>TG con-MNV$_{1\text{bp}}$s. This was also the case for 96% and 92% of the other less common possible CpG creating con-MNVs CC>TG and AG>CA. CA>TG is probably the most common variant as it relies on a transition mutation A>G happening first which has a higher mutation rate compared to the transversions C>G and T>G. I identified 255 *de novo* con-MNVs and 26 of these were *de novo* con-MNV$_{1\text{bp}}$s. In half of these, the inherited variant created a CpG site which was then mutated *de novo* in our data.

I also observed that for con-MNV$_{3\text{bp}}$s that were not as a result of CpG creating sites, the first variant increases the mutability of the second variant more than expected by chance. I compared the median difference in mutation probability of the second variant based on the heptanucleotide sequence context before and after the first variant occurred using a signed Wilcoxon Rank Test[3]. The median increase in mutation probability of the second variant was 0.0002 (p = 9.8×10^{-17},signed Wilcoxon rank test).

2.3.3 Misannotation of MNVs

When an MNV occurs within a single codon, the consequence of this MNV can be different to the consequences when the two comprising variants are annotated separately. For 98% of the intra-codon MNVs, the consequence class (synonymous, missense, stop-gained etc.) of the MNV was the same as at least one of the SNVs annotated separately. For only 1% of the intra-codon MNVs was the consequence class of the MNV more severe than the separate SNVs. For almost all of these the MNV caused a stop-gain. Most intra-codon MNVs result in a missense change (Table 2.2) and so even though one of the comprising variants is most likely annotated as a missense separately as well, the MNV can create a different amino acid change.

MNV Consequence	Sim- MNV (% of all sim-MNVs)	Con-MNV (% of all con-MNVs)
Synonymous	10 (0.5%)	5 (0.3%)
1-step missense	815 (38.2%)	814 (50.7%)
2-step missense	1265 (59.2%)	757 (47.2%)
Stop Loss	2 (0.1%)	4 (0.2%)
Stop Gain	44 (2.0%)	24 (1.5%)

Table 2.2 Numbers and proportions of consequence types for MNVs within same codon

2.3.4 Functional Consequences of MNVs

The structure of the genetic code is not random. The code has evolved such that the codons that correspond to amino acids with similar physiochemical properties are more likely to be separated by a single base change [7, 233]. SNVs that result in a missense change will only alter one of the bases in a codon, however MNVs that alter a single codon (‘intra-codon’ MNVs) will alter two of the three base pairs. Therefore, they are more likely to introduce an amino acid that is further away in the codon table and thus less similar physicochemically to the original amino acid. Most intra-codon MNVs result in a missense change (Table 2.2). Intra-codon missense MNVs can be classified into two groups: ‘one-step’ and ‘two-step’ missense MNVs. One-step missense MNVs lead to an amino acid change that could also have been achieved by an SNV, whereas two-step MNVs generate amino-acid changes that could only be achieved by two SNVs. For example if we consider the codon CAC which codes for Histidine (H) then a single base change in the codon can lead to missense changes creating seven possible amino acids (Y,R,N,D,P,L,Q) (Figure 2.5a). There are one-step missense MNVs within that codon that can lead to most of the same amino acids (Y,R,N,D,P,L). However two-step missense MNVs could also lead to an additional eleven amino acids that could not be achieved by an SNV (F,S,C,I,T,K,S,V,A,E,G). For some codons there are also amino acid changes that can only be created by a single base change, for this Histidine

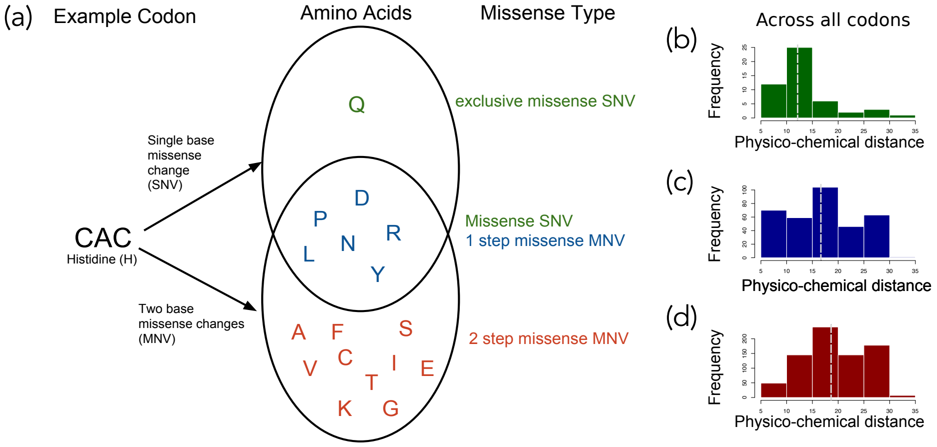


Fig. 2.5 Classification of intra-codon MNV missense mutations (a) Example of how one-step missense MNVs and two-step missense MNVs are classified using a single codon ’CAC’. Venn diagram shows amino acids that can be created with either a single base change or a two base change in the codon ’CAC’.(b-d) Across all codons the distribution of physiochemical distances for the amino acid changes caused by different types of missense variants, dashed line indicates the median of the distribution (b) exclusive SNV missense (c) one-step MNV missense (d) two-step MNV missense

codon this would be Glutamine (Q). These will be referred to as exclusive SNV missense changes. For this analysis I only considered sim-MNVs that most likely originated from the same mutational event. This is because I was primarily interested in the functional effects of mutations occurring simultaneously and where the amino acid produced would have changed directly from the original amino acid to the MNV consequence and not via an intermediate amino acid.

2.3.5 MNVs can create a missense change with a larger physico-chemical distance compared to missense SNVs

I assessed the differences in the amino acid changes between exclusive missense SNVs, one-step MNVs and two-step MNVs by examining the distribution of physicochemical distance for each missense variant type across all codons (Figure 2.5b). I used a distance measure between quantitative descriptors of amino acids based on multidimensional scaling of 237 physical-chemical properties[223]. I chose this measure as it does not depend on observed substitution frequencies which may create a bias due to the low MNV mutation rate making these amino acid changes inherently less likely. The median amino acid distance was significantly larger for two-step missense MNVs when compared to one-step missense MNVs ($p = 1.10 \times 10^{-7}$, Wilcoxon test). The median distance for one-step missense MNVs was also significantly larger from exclusive SNV missense changes (p = 0.0008, Wilcoxon test) (Figure 2.5b-d).

2.3.6 Missense MNVs are on average more damaging than missense SNVs

If the physico-chemical differences between these classes of missense variants resulted in more damaging mutations in the context of the protein, then I would expect to see a greater depletion of two-step missense MNVs compared to one-step missense MNVs or missense SNVs in highly constrained genes. I looked at the proportion of variants of different classes that fell in highly constrained genes, as defined by their intolerance of truncating variants in population variation, as measured by the probability of loss-of-function intolerance (pLI) score (Figure 2.6a). Highly constrained genes were defined as those with a pLI score >=0.9 [125]. MNVs that impact two nearby codons (inter-codon MNVs) are likely to have a more severe consequence on protein function, on average, than an SNV impacting on a single codon. I observed that the proportion of inter-codon $\text{MNV}_{\text{1-20bp}}$s that fall in highly constrained genes (pLI>0.9) is significantly smaller compared to missense SNVs (p = 0.0007, proportion test)

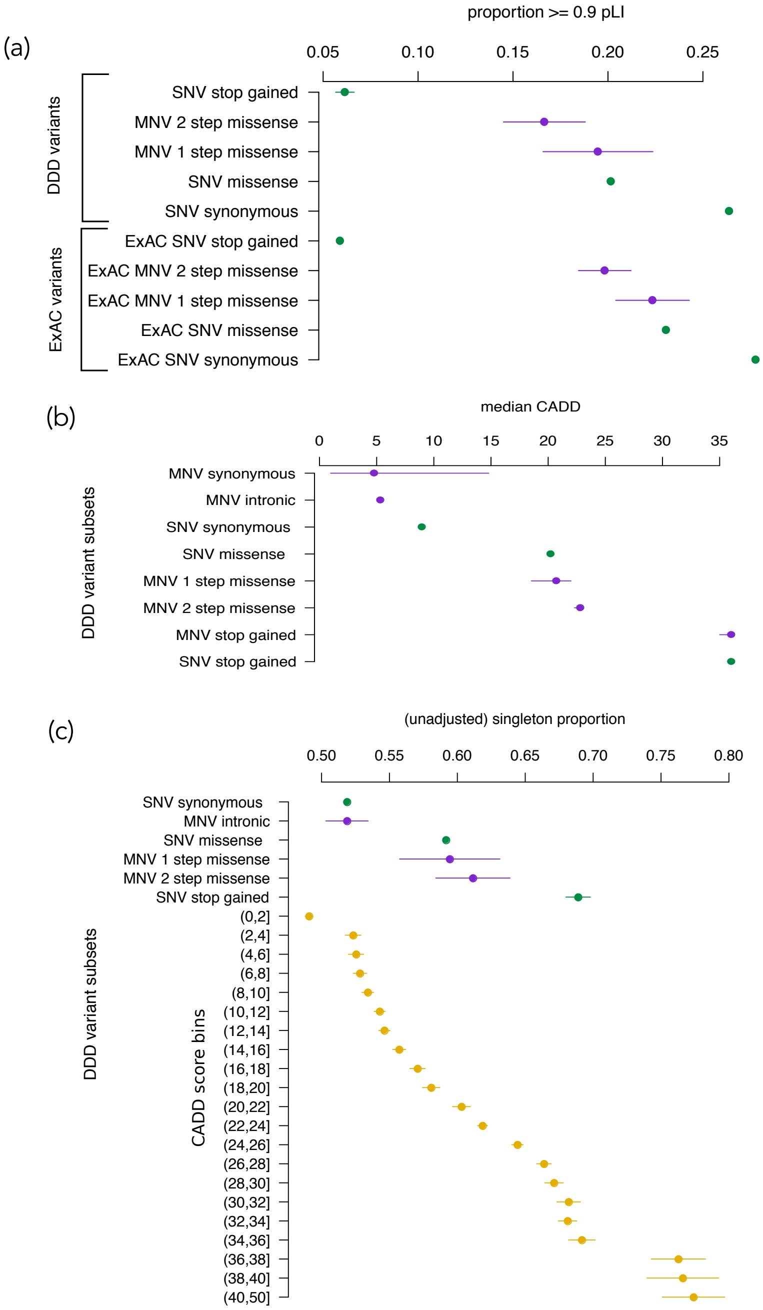


Fig. 2.6 Quantifying the pathogenicity of MNVs.(a) Proportion of variants that fall in genes with pLI >= 0.9 over different classes of variants for both DDD and ExAC datasets. Green are SNVs, Purple are MNVs. Lines are 95% confidence intervals.(b) The median CADD score over different classes of variants identified from DDD data with bootstrapped 95% confidence intervals.(c) Singleton proportion for different classes of DDD variants. In yellow are SNVs stratified by binned CADD scores with their corresponding singleton proportions. Lines are 95% confidence intervals.

(Figure 2.6a). For intra-codon MNVs, the proportion of two-step missense MNVs observed in highly constrained genes was also significantly smaller than for missense SNVs (p = 0.0016, Proportion test). The proportion of one-step missense MNVs was not significantly different from either missense SNVs or two-step missense MNVs. The analysis was repeated using SNVs and MNVs that were identified by the Exome Aggregation Consortium (ExAC) that were subject to different filtering steps [125]. The same relationship was observed, the proportion of ExAC two-step MNVs in high pLI genes was significantly smaller than for ExAC missense SNVs ($p = 9.84 \times 10^{-6}$).

I then compared variant deleteriousness across the variant classes using Combined Annotation Dependent Depletion (CADD) score that integrates many annotations such as likely protein consequence, constraint and mappability[107](Figure 2.6b). I found that the median CADD score for two-step missense MNVs was significantly higher than both one-step missense MNVs (p = 0.00017, Wilcoxon test) and missense SNVs ($p = 2.70 \times 10^{-8}$, Wilcoxon test). Two-step MNV missense had a median CADD score of 22.8 compared to a one-step missense median CADD score of 20.7 and a SNV missense median CADD score of 20.2.

The proportion of singletons across variant classes is a good proxy for the strength of purifying selection acting in a population[125]. The more deleterious a variant class is, the larger the proportion of singletons. We found that the singleton proportion for two-step missense MNVs was nominally significantly higher compared to missense SNVs (p = 0.02, proportion test) (Figure 2.6c). This increase in singleton proportion corresponded to an increase of about two in the interpolated CADD score. This is concordant with the increase in CADD scores that was computed directly above.

2.3.7 Estimation of the MNV mutation rate

I estimated the genome-wide mutation rate of sim-$\text{MNV}_{\text{1-20bp}}$s to be 1.78×10^{-10} mutations per base pair per generation by scaling the SNV mutation rate based on the relative ratio of segregating polymorphisms for MNVs and SNVs [226], see Methods. For this estimate I only used variants that fell into non-constrained genes (pLI<0.1) and non-coding regions to avoid any bias from ascertainment bias. I assumed that recurrent mutation is insufficiently frequent for both classes of variation to alter the proportionality between the number of segregating polymorphisms and the mutation rate. This estimate is ~1.6% the mutation rate estimate for SNVs and accords with the genome-wide proportions of SNVs and MNVs described previously [191]. I was concerned that the selective pressure on MNVs and SNVs might still be different in non-constrained genes and this could affect the mutation rate estimate. To see if this was the case, I applied the same method to estimate the SNV missense mutation rate across coding region and found that the estimate was concordant with that obtained from using an SNV tri-nucleotide context mutational model[187]. I also estimated the MNV mutation rate using the set of *de novo* MNVs that fell into non-constrained genes (pLI<0.1) that have not previously been associated with dominant developmental disorders and obtained a concordant mutation rate estimate of 1.79×10^{-10} (confidence interval: 0.88×10^{-10}, 2.70×10^{-10}) mutations per base pair per generation, very similar to the estimate based on segregating polymorphisms described above.

2.3.8 Contribution of *de novo* MNVs to developmental disorders

I identified 10 *de novo* MNVs within genes known to be associated with dominant developmental disorders (DD-associated) in the DDD trios (Table 2.3), which is a significant (p = 1.03×10^{-3},Poisson test) 3.7 fold enrichment compared with what we would expect based on our estimated MNV mutation rate. This enrichment is similar in magnitude to that observed for *de novo* SNVs in the same set of DD-associated genes (Figure 2.8).

Decipher ID	Distance between variants	Chr	Positions	Gene	Ref	Alt	Consequence (first variant/ second variant)	MNV falls within/between codon	Clinician pathogenicity annotation on Decipher
261423	1	5	161569244, 161569245	*GABRG2*	CC	TT	missense (two step)	Within codon	Likely pathogenic (Full)
292136	1	14	29237129, 29237130	*FOXG1*	TC	CT	missense (one step)	Within codon	Likely pathogenic (Full)
280956	1	19	13135878, 13135879	*NFIX*	GC	TT	missense (one step)	Within codon	Likely pathogenic (Partial)
270803	1	3	49114312, 49114313	*QRICH1*	GC	AA	stop gain/missense	Between codon	Likely pathogenic (Partial)
258688	1	5	67591021, 67591022	*PIK3R1*	TA	GC	missense/missense	Between codon	Likely pathogenic (Full)
274482	1	16	30749053, 30749054	*SRCAP*	GG	AT	synonymous/ stop gain	Between codon	Definitely pathogenic (Full)
274606	1	9	140637863, 140637864	*EHMT1*	GA	TT	missense/stop gain	Between codon	Likely pathogenic (Full)
274453	1	9	140637863, 140637864	*EHMT1*	GA	TT	missense/stop gain	Between codon	Definitely pathogenic (Full)
260753	13	6	157454286, 157454297	*ARID1B*	G..C	T..G	missense/stop gain	Between codon	Definitely pathogenic (Full)
270916	3	1	7309651, 7309654	*CAMTA1*	G..G	A..A	missense/missense	Between codon	Likely pathogenic (partial)

Table 2.3 *De Novo* MNVs that fall in genes associated with developmental disorders

To assess the sensitivity of this enrichment to the estimate of the MNV mutation rate I recalculated this by using an MNV mutation rate estimate based on all variants, as opposed to excluding those that fall in DDG2P genes, as well as a more stringent estimate just using variants that fell into non-coding regions. When I redid the enrichment analysis using these mutation rate estimates of varying stringency, the enrichment of *de novo* MNVs in DD-associated genes remained significant (all variants p = 2.7×10^{-4}, non coding control regions p = 4.9×10^{-3}, Figure 2.7a). The SNV mutation rate estimate varies across studies therefore I also recalculated the MNV mutation rates using SNV mutation rate estimates of 1.0×10^{-8} and 1.2×10^{-8} mutations per base pair per generation [195]. These were also recalculated across the three different variant subsets (all variants, excluding variants in genes with pLI>0.1, variants in non-coding control regions). The enrichment ratio of *de novo* MNVs that fall into DD genes ranged from 2.7 to 4.8 however always remained significantly greater than 1 and the confidence intervals consistently overlapped with that of the SNV missense enrichment ratio (Figure 2.7b).

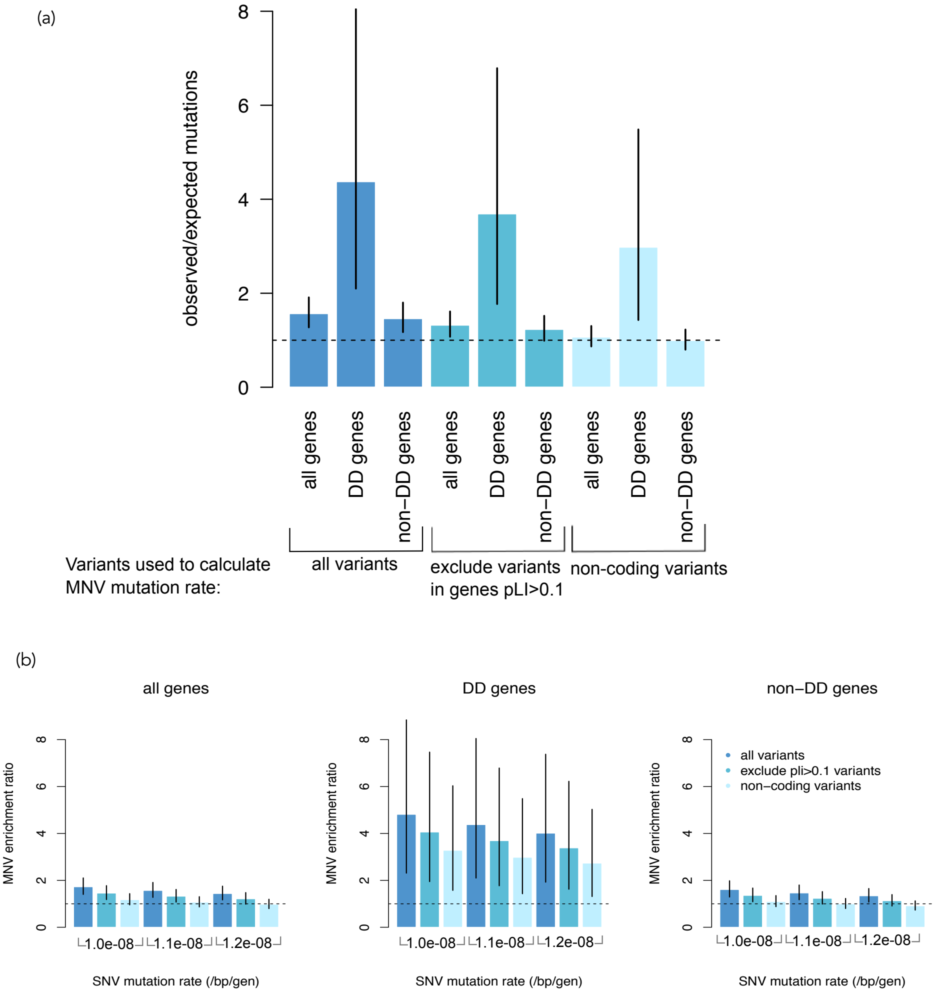


Fig. 2.7 Sensitivity of MNV enrichment analysis to MNV mutation rate estimates (a) The impact of varying the subsets of variants used to estimate the MNV mutation rate estimate on the enrichment of *de novo* MNVs in different subcategories of genes as in Figure 2.8. These were all calculated using an SNV mutation rate estimate of 1.1×10^{-8} /bp/generation. (b) Using three different estimates of the SNV mutation rate estimate and the subcategories of variants as in (a) looking at the difference in enrichment ratios across the same subcategories of genes as in (a).

To ensure this observed enrichment was not driven by differences in sequence context, I also evaluated whether DD-associated genes are enriched for the primary mutagenic dinucleotide contexts associated with the signatures of polymerase ζ. I found that DD-associated genes had a small (1.02 fold) but significant ($p = 1.9 \times 10^{-59}$,proportion test) enrichment of polymerase ζ dinucleotide contexts compared to genes not associated with DD. However, this subtle enrichment is insufficient to explain the four-fold enrichment of *de novo* MNVs in these genes. The enrichment for *de novo* MNVs remains significant after correcting for this sequence context ($p = 2.28 \times 10^{-3}$, Poisson test).

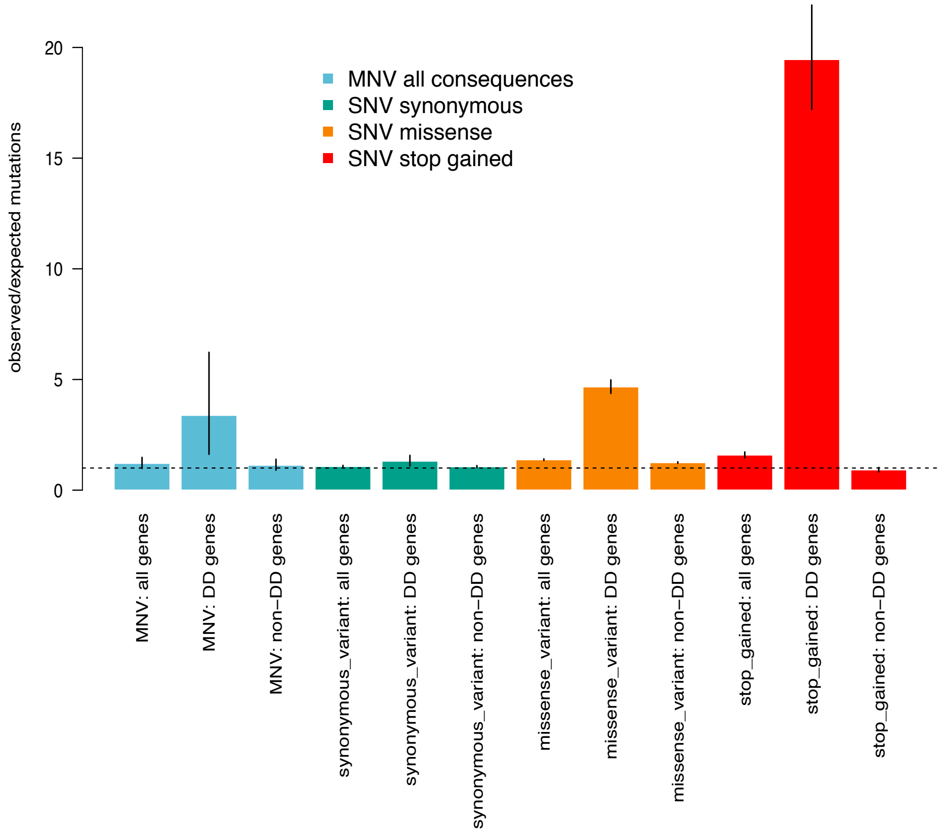


Fig. 2.8 Enrichment of *de novo* MNVs in DDD study. Ratio of observed number of *de novo* MNVs vs the expected number of *de novo* MNVs based on the estimate of the MNV mutation rate. Compared to enrichment of SNVs in DD genes in consequence classes synonymous, missense and stop gain.

Eight of the 10 *de novo* MNVs in DD-associated genes were 1 bp apart while the other two were 3 and 13 bp apart. All of these *de novo* MNVs were experimentally validated in the child (and their absence confirmed in both parents) using either MiSeq or capillary sequencing. This validation was done by Elena Prigmore. All ten MNVs were thought to be pathogenic by the child's referring clinical geneticist. Seven of the MNVs impacted two different codons while three fell within the same codon, one of which created a two-step missense change. Of those MNVs that impacted two codons, five caused a premature stop codon. I found a recurrent *de novo* MNV in the gene *EHMT1* in two unrelated patients that bore the distinctive polymerase ζ signature of GA>TT.

2.3.9 Clinically reported MNVs in DD-associated genes

To assess whether MNVs are being underreported in genes associated with DD, I downloaded all clinically reported variants in DD-associated genes from ClinVar (accessed September 2017, [115]). I looked at the number of intra-codon missense MNVs in genes that have at least one reported pathogenic missense mutation. This was to ensure that missense mutations in that gene would likely cause DD. I focused on intra-codon MNVs as it is the interpretation of this class of MNV that is most impacted by failing to consider the variant as single unit. I calculated the expected number of pathogenic MNVs in these genes based on the MNV mutation rate and the number of pathogenic SNV missense variants reported. There were 22 reported pathogenic MNVs compared to the expected number of 26 across 321 genes which was not significantly different (p = 0.55, Poisson test). I also looked for clinically relevant SNVs in ClinVar that overlapped with inherited sim-MNVs that I identified in our data. I found one SNV that had been reported as a nonsense variant in the gene *AGPAT2*. The variant had been reported as pathogenic and of uncertain significance for congenital generalised lipdystrophy type 1 by two contributors. However I observe this variant as an MNV in our dataset in three individuals. The MNV falls within the same codon and causes a missense as opposed to a stop gain which decreases its likelihood of pathogenicity, especially since it was also observed 70 times in ExAC (Allele Frequency 3.96^{-4}).

2.3.10 MNV mutator phenotype

Five DDD children had more than one *de novo* sim-MNVs. This is significantly greater than what we would expect assuming these MNVs arose independently. Using our estimated MNV mutation rate, the probability of seeing five or more individuals in our data set with more than one MNV is 5.8×10^{-7}. The number of MNVs per person range from 2-5 *de novo* MNVs. These mostly appear on different chromosomes and have different distances between the pair of variants. None of the MNVs share the same mutation and the number of mutations is too small to pick up on more subtle similarities in the mutational signatures. A comparable mutator phenotype has been observed in other classes of genetic variation such as CNVs but, similarly, a relevant mutational mechanism has not yet been discovered [131]. A larger number of *de novo* MNVs will help to uncover possible mechanisms behind this apparent mutator phenotype.

2.4 Discussion

MNVs constitute a unique class of variant, both in terms of mutational mechanism and functional impact. I found that 18% of segregating MNVs were at adjacent nucleotides and estimated that 19% of all MNVs represent a single mutational event, increasing to 53% of $\mathrm{MNV}_{1\mathrm{bp}}$. I estimated the sim-MNV germline mutation rate to be 1.78×10^{-10} mutations per base pair per generation, roughly 1.6% that of SNVs. Most population genetics models assume that mutations arise from independent events [76]. MNVs violate that assumption and this may affect the accuracy of these models. Recent studies suggest that certain phylogenetic tests of adaptive evolution incorrectly identify positive selection when the presence of these clustered mutations are ignored [222]. Correcting these population genetic models will require knowledge of the rate and spectrum of MNV mutations. The observation of a possible MNV mutator phenotype complicates this correction further. In the future it would be of interest to whole-genome sequence those individuals with a potential MNV mutator phenotype to uncover the causal underlying mechanisms. I replicated the observations from previous studies that several different mutational processes underlie MNV formation, and that these tend to create MNVs of different types. Error-prone polymerase ζ predominantly creates sim-$\mathrm{MNV}_{1\mathrm{bp}}$ [76, 14]. APOBEC-related mutation processes have been described to generate MNVs in the range of 10-50bp [181, 5, 77], but here I show that an enrichment for APOBEC motifs can be detected down to $\mathrm{MNV}_{2\mathrm{bp}}$. Nonetheless, there remain other sim-MNVs that cannot be readily explained by either of these mechanisms, and it is likely that other, less distinctive, mutational mechanisms remain to be delineated as catalogs of MNVs increase in scale. These future studies should also investigate whether these MNV mutational signatures differ subtly between human populations as has been recently observed for SNVs [75]. Consecutive MNVs, by contrast, exhibit greater similarity with known SNV mutation processes, most notably with the creation and subsequent mutation of mutagenic CpG dinucleotides. The non-Markovian nature of this consecutive mutation process challenges Markovian assumptions that are prevalent within standard population genetic models [179].

These findings validated the intuitive hypothesis that MNVs that impact upon two codons within a protein are likely, on average, to have a greater functional impact than SNVs that alter a single codon. I evaluated the functional impact of intra-codon MNVs using three complementary approaches: (i) depletion within genes under strong selective constraint, (ii) shift towards rarer alleles in the site frequency spectrum and (iii) enrichment of *de novo* mutations in known DD-associated genes in children with DDs. I demonstrated that intra-codon MNVs also tend to have a larger functional impact than SNVs, and that MNV missense changes that cannot be achieved by a single SNV are, on average, more deleterious than those that can. This is most likely due to the fact that they are on average more physico-chemically different compared to amino acids created by SNVs and are not as well tolerated in the context of the encoded protein. These ‘two-step’ missense MNVs make up more than half of all sim-MNVs that alter a single codon. I also identified 10 pathogenic *de novo* MNVs within the DDD study, including both intra-codon and inter-codon MNVs. With larger trio datasets we will have more power to tease apart more subtle differences in pathogenic burden and purifying selection between different classes of MNVs and SNVs, for example, to test whether two-step missense *de novo* MNVs are more enriched than missense SNVs or one-step missense MNVs in developmental disorders. More data will also allow us to assess the population genetic properties of inter-codon MNVs.

These findings emphasise the critical importance of accurately calling and annotating MNVs within clinical genomic testing both to improve diagnostic sensitivity and to avoid misinterpretation. While MNVs are not underrepresented in reported clinically reported variants in ClinVar, we did observe that pathogenic *de novo* MNVs can be mis-annotated, indicating that current analytical workflows may not be calling these correctly. In a recent comparison of eight different variant calling tools it was noted that only two callers, FreeBayes and VarDict, report two mutations in close proximity as MNVs. The others reported them as two separate SNVs [188]. Both FreeBayes and VarDict are haplotype aware callers which is necessary for MNV detection [56, 114]. Even if variant callers do not identify MNVs directly, software also exists that can correct a list of previously called SNVs to identify mis-annotated MNVs [228]. To further our understanding of the role of MNVs in evolution and disease, calling and annotating these variants correctly is a vital step.

Chapter 3

Identifying and characterising germline hypermutators

3.1 Introduction

Germline mutagenesis is a major source of all genetic variation and drives the process of human evolution. The human mutation rate is not a constant, the rate of mutation varies both within and between individuals. For example, parental age explains a large proportion of variance between individuals [111, 60]. The factors influencing variation in germline mutation rates are discussed extensively in Chapter 1. While we have started to explain the general distribution of mutations, little is known about rare outliers with extreme mutation rates. Germline hypermutators are defined here as individuals with an unusually large germline mutation rate. This may be due to environmental factors or could have a genetic basis.

The impact of environmental mutagens has been well established in the soma but this is not as well understood in the germline. Environmental exposures in parents can influence the number of mutations transmitted to offspring. For example, ionising radiation has a mutagenic effect on the germline and offspring of irradiated parents are observed to have an increased number of *de novo* mutations [213]. By comparing nearby populations in regions with differing levels of natural radiation, it has also been observed that radiation increases the rate of mutation in mitochondrial DNA[51]. Exposure to ionising radiation has been confirmed to increase the paternal germline mutation rate in mice *in vivo* [2]. Tobacco smoke has also been hypothesised to increase the number of mutations in the paternal germline. Exposure to tobacco smoke has been observed to increase the number of mutations, specifically at short tandem repeats, in spermatogonial stem cells in mice [242, 139, 11].

Individual mutation rate can also be influenced by genetic background. With regards to somatic mutation, thousands of inherited germline variants have been shown to predispose individual cancer risk [82, 46, 81]. For example, Li-Fraumeni syndrome (LFS) is an autosomal dominant disorder which leads to a large increased risk of early-onset cancers due to inherited germline variants in the gene *TP53*. This elevated rate of mutation, and resulting increased cancer risk, also appears to extend to the germline as families with LFS are also highly enriched for *de novo* CNVs compared to the healthy population [200]. Homozygous and heterozygous germline PTVs in *NTHL1*, a gene involved in the base excision repair pathway, are another example; these variants have been shown to predispose individuals to colorectal cancer[231]. This raises the question of whether other known cancer predisposing variants also impact the rate of mutation in the germline. Pathogenic variants in the gene *MBD4* have been shown to elevate cancer risk, primarily for colorectal cancers. A recent study investigating genetic determinants of cancer identified that patients with germline heterozygous protein truncating variants (PTVs) in the *MBD4* gene have a four-fold increase in C>T mutations at CpG dinucleotides in their tumours [214]. This result agrees with previous studies that showed that Mbd4 knockout (*Mbd4* -/-) in mice was found to accelerate tumorigenesis and mutation analysis of these tumours showed a three-fold increase in the number of C>T mutations at CpGs [147, 232]. *MBD4* is known to play a role in base-excision repair. Specifically it encodes a DNA glycolysase that removes thymidines from T:G mismatches at methyl-CpG sites [79]. Many of these variants associated with elevated cancer risk are in genes encoding components of DNA repair pathways which, when impaired, lead to an increased number of somatic mutations. However it is not known whether variants in known somatic mutator genes can influence germline mutation rates. For example, the CpG mutation signature (Signature 1 in the catalogue of somatic mutations in cancer (COSMIC)) accounts for 16% of *de novo* mutations in the germline which raises the question of whether *MBD*4 PTV germline carriers also show an increased number of C>T germline mutations in their offspring.

There have been several examples where genetic background has been shown to impact the germline mutation rate of a variety of types of genetic variation. As mentioned in Chapter 1, the mutation rate of STRs are known to be affected by both the length of the repeat unit and the repeat number [84, 70, 210]. Variants have also been shown to impact the mutation rate of minisatellites. Through the analysis of single sperm, a variant nearby to minisatellite MS32 has been shown to impact its mutation rate in the male germline[150]. With respect to translocations, an analysis of a recurrent chromosomal translocation demonstrated that the breakpoints occur in the center of a region of palindromic AT-rich repeats (PATRRs). The presence of PATRR-like sequences was also identified at other translocation breakpoints which suggests that these regions are susceptible to double strand breaks and likely increase the rate of translocation[103]. A recent study has also attempted to identify variants associated with overall germline mutation rate by leveraging a haplotype based approach[197]. This was based on the idea that when a variant increases the germline mutation rate it results in a subset of haplotypes that are more divergent than others at that locus. With this method, the authors identified several candidate mutator loci and found these were enriched for their proximity to genes associated with DNA repair.

An elevated germline mutation rate can have a significant impact on the health of subsequent generations. Increasing germline mutation rate results in an increased risk of offspring being born with a congenital disorder caused by a *de novo* mutation. There are also long-term effects of mutation rate differences. The phenotypic effects of mutation accumulation were examined in homozygous *Pold1* knockout mice[219]. These mice have an ~17 fold increased germline mutation rate due to the lack of the proofreading activity of DNA polymerase delta. Abnormal phenotypes were observed ~4 times more than in controls and after several generations the *Pold1* deficient mice had much lower reproduction rates with lower pregnancy rates, lower survival rates and smaller litter sizes[219]. A recent study examined a set of 41 multi-generational families and observed that a higher germline mutation rate is correlated with higher all-cause mortality and reduced fertility in women [25]. The decrease in fertility is suggested to be due to germline mutation accumulation while the shorter lifespan is hypothesised to be driven by a correlation between germline and somatic mutation rates.

3.1.1 Chapter Overview

In this chapter, I used large cohorts of exome and whole-genome sequenced parent-offspring trios in order to investigate germline hypermutators and the impact of rare genetic variation on individual mutation rates. I focussed on SNV mutation rates specifically in this chapter and tackled this problem from two different angles: a genotype-driven approach and subsequently a phenotype-driven approach.

The genotype-driven approach focused on whether variants in DNA repair genes impact germline mutation rates. For this, I interrogated variants in an established cancer mutator gene, *MBD4*, to investigate if they have a similar effect in the germline.

For the phenotype-driven approach, I aimed to identify germline hypermutators and sought genetic causes for this trait. Germline hypermutators are individuals who have an elevated germline mutation rate and so are likely to have children with an unusually large number of *de novo* mutations (DNMs). A large number of DNMs increases the chance of having a dominant disorder, therefore cohorts of children with rare disease are better powered to identify germline hypermutators. A CNV mutator phenotype has been previously identified in a cohort of patients with neurodevelopmental phenotypes [131]. I then identified the fraction of variance in germline mutation rate that can be explained by parental age and hypermutation and explore where the remaining fraction of variance may lie by performing analyses of the impact of rare damaging variants in DNA repair genes on an individual's germline mutation rate across the 100kGP dataset.

In this chapter I also take advantage of the large number of WGS trios available in the 100,000 Genomes Project (100kGP) to examine other sources of mutation rate variability such as the effects of parental age.

3.1.2 Contributions

I would like to acknowledge Jan Korbel who initially approached us regarding the possibility that *MBD4* may impact the germline. The *de novo* filtering for the Genomics England dataset was done in collaboration with Patrick Short, Chris Odhams and Loukas Moutsianas. Petr Danecek collated the variants in DNA repair genes in the 100kGP dataset which I used for the analysis looking at the impact of these variants on germline mutation rate. James Stephenson, a post-doc in the group, helped analyse the role of a variant within the context of the protein structure. Raheleh Rahbari and Matthew Neville advised on the mutational signature extraction. This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. All of this work was done under the supervision of Matthew E. Hurles

3.2 Methods

3.2.1 *De novo* calling and filtering in paternal *MBD4* PTV carriers

I identified 14 individuals in the DDD study whose father had a heterozyous protein-truncating variant (PTV) in the gene *MBD4*. This included five stop gained variants, 7 frameshift variants and 2 variants within splice donor sites. There were 11 unique variant sites. All variants were examined in the Integrative Genomics Viewer (IGV) and did not appear to be false positive sites. These 14 parent offspring trios (42 samples) were submitted for whole-genome sequencing PCR-free at >30x mean coverage of Illumina 150 bp paired end reads via Sanger pipelines. One sample failed at the library creation phase and there was not enough sample left to resubmit. This left me with 13 trios for analysis. The reads were mapped with bwa (v0.7.16). I used GATK (v3.5) HaplotypeCaller best practices to generate a multi-sample VCF and from this created parent-offspring trio VCFs and the input files needed for DNM calling. DNMs were called in these trios using bcftool's trio-dnm. This was a change from how DNMs were called previously in DDD using DeNovoGear (the DNM caller described in Chapter 2 and 4), which no longer functioned efficiently on the Sanger compute cluster. The filters selected here were chosen after inspecting distributions of variant allele fraction (VAF) and examination of these putative DNMs using the Integrative Genomics Viewer (IGV) to estimate true positive rates.

Filters applied:

- Removed DNMs with trio-dnm 'DNM' score < 50, this is the score outputted by trio-dnm. It is the log of the probability of inheriting the variant calculated directly from the genotype likelihoods.
- Removed DNMs that fell within known segmental duplication regions as defined by UCSC (http://humanparalogy.gs.washington.edu/build37/data/GRCh37GenomicSuperDup.tab)
- Removed DNMs that fell in highly repetitive regions (http://humanparalogy.gs.washington.edu/build37/
- Removed DNMs with gnomAD allele frequency > 0.01
- Read depth (RD) of child > 7, mother RD > 5, father RD > 5
- Alternative allele depth of child >2
- Fisher exact test on strand bias p-value > 10^{-3}
- Removed DNMs if child RD >98 [173]
- Remove DNMs with >1 alternative read in either parent
- Remove DNMs with > 0.1 parental VAF in either parent
- Test to see if VAF in child is significantly greater than the error rate at that site as defined by error sites estimated using Shearwater. This was calculated by Inigo Martincorena [57].

This resulted in 1,690 DNMs after this stage of filtering (~130 per trio). I examined all of these with IGV and annotated them with whether these appeared true. This resulted in a total of 877 DNMs across the 13 trios (~67 DNMs per person). Due to the small number of trios I examined here I did not refine my filters based on this annotation but plan to do so in order to improve DNM calling with bcftools for DDD trios in the future.

3.2.2 DNM filtering in 100,000 Genomes Project

I analysed DNMs called in 13,949 parent offspring trios from 12,609 families from the rare disease programme. Sequencing and variant calling for these families was performed via the Genomics England rare disease analysis pipeline which has been extensively documented (https://cnfl.extge.co.uk/display/GERE/10.+Further+reading+and+documentation). DNMs were called by the Genomics England Bioinformatics team using the Platypus variant caller[178]. Filtering of the DNMs was done in collaboration with Patrick Short, Chris Odhams and Loukas Moutsianas. These were selected to optimise various properties including the number of DNMs per person being approximately what we would expect, the distribution of the VAF of the DNMs to be centered around 0.5 and the true positive rate of DNMs to be sufficiently high as calculated from examining IGV plots. The filters applied were as follows:

- Genotype is heterozygous in child (1/0) and homozygous in both parents (0/0)
- Child RD >20, Mother RD>20, Father RD>20
- Remove variants with >1 alternative read in either parent
- VAF>0.3 and VAF<0.7 for child
- Remove SNVs within 20 bp of each other. While this is likely removing true MNVs, the error mode was very high for clustered mutations.
- Removed DNMs if child RD >98 [173]
- Removed DNMs that fell within known segmental duplication regions as defined by UCSC (http://humanparalogy.gs.washington.edu/build37/data/GRCh37GenomicSuperDup.tab)
- Removed DNMs that fell in highly repetitive regions (http://humanparalogy.gs.washington.edu/build37/dat
- For DNM calls that fell on the X chromosome these slightly modified filters were used:
  - For DNMs that fell in PAR regions, the filters were unchanged from the autosomal calls apart from allowing for both heterozygous (1/0) and hemizygous (1) calls in males
  - For DNMs that fell in non-PAR regions the following filters were used:

  - ∗ For males: RD>20 in child, RD>20 in mother, no RD filter on father
  - ∗ For males: the genotype must be hemizygous (1) in child and homozygous in mother (0/0)
  - ∗ For females: RD>20 in child, RD>20 in mother, RD>10 in father

3.2.3 DNM filtering for possible DDD hypermutated individuals

Nine trios were selected from the DDD cohort where the offspring has an unusually large number of exome DNMs and submitted along with their parents for whole-genome sequencing PCR-free at >30x mean coverage of Illumina 150bp paired end reads via Sanger pipelines. Reads were mapped with bwa (v0.7.15). DNMs were called from these trios using DeNovoGear[174] (note this analysis was done over a year prior to the *MBD4* analysis which is why DeNovoGear was used here) and were filtered as follows:

- Read depth (RD) of child > 10, mother RD > 10, father RD > 10
- Alternative allele read depth in child >2
- Filtered on strand bias across parents and child (p-value >0.001, Fisher's exact test)
- Removed DNMs that fell within known segmental duplication regions as defined by UCSC (http://humanparalogy.gs.washington.edu/build37/data/GRCh37GenomicSuperDup.tab)
- Removed DNMs that fell in highly repetitive regions (http://humanparalogy.gs.washington.edu/build37/
- Allele frequency in gnomAD < 0.01
- VAF <0.1 for both parents
- Removed mutations if both parents have >1 read supporting the alternative allele
- Test to see if VAF in child is significantly greater than the error rate at that site as defined by error sites estimated using Shearwater. This was calculated by Inigo Martincorena [57].
- Posterior probability from DeNovoGear > 0.00781 [41].
- Removed DNMs if child RD >200 [173].

After applying these filters, this resulted in 1,367 DNMs. I then inspected all of these DNMs using IGV and removed those that appeared to be false positives. I had a final set of 916 DNMs across the 10 trios.

3.2.4 Parental phasing of *de novo* mutations

To phase the DNMs in both 100kGP and DDD I used a custom script which used the following read-based approach to phase a DNM. I first searched for heterozygous variants within 500 bp of the DNM that was able to be phased to a parent (so not heterozygous in both parents and offspring). I then examined the reads or read pairs which included both the variant and the DNM and counted how many times I observe the DNM on the same haplotype of each parent. If the DNM appears exclusively on the same haplotype as a single parent then that was determined to originate from that parent. I discarded DNMs that had conflicting evidence from both parents. This code is available on GitHub (https://github.com/queenjobo/PhaseMyDeNovo).

3.2.5 Analysis of effect of parental age on germline mutation rate

To assess the effect of parental age on germline mutation rate I ran the following regressions. On all (unphased) DNMs I ran two separate regressions for SNVs and indels. I fitted the following model using a Poisson generalized linear model (GLM) with an identity link where Y is the number of DNMs for an individual:

$$E(Y) = \beta_0 + \beta_1 paternal\_age + \beta_2 maternal\_age \tag{3.1}$$

For the phased DNMs I fit the following two models using a Poisson GLM with an identity link where $Y_{maternal}$ is the number of maternally derived DNMs and $Y_{paternal}$ is the number of paternally derived DNMs:

$$E(Y_{paternal}) = \beta_0 + \beta_1 paternal\_age$$
$$E(Y_{maternal}) = \beta_0 + \beta_1 maternal\_age$$

3.2.6 Identifying hypermutation in 100kGP

To identify hypermutated individuals in the 100kGP cohort I first wanted to regress out the effect of parental age by fitting the following ordinary linear regression model:

$$E(Y) = \beta_0 + \beta_1 paternal\_age + \beta_2 maternal\_age \tag{3.2}$$

I then looked at the distribution of the studentized residuals and then, assuming these followed a t distribution with N-2-1 degrees of freedon, calculated a t-test p-value for each individual. I separately did the same approach for the number of indels, except in this case Y would be the number of *de novo* indels.

3.2.7 Extraction of mutational signatures

I extracted mutational signatures from maternally and paternally phased DNMs as well as from the 15 hypermutated individuals that I identified. I did this using SigProfiler (v1.0.5) and these signatures are extracted and subsequently mapped on to COSMIC mutational signatures (COMIC v89, Mutational Signature v3) [12, 212].

3.2.8 Defining set of genes involved in DNA repair

I compiled a list of DNA repair genes which were taken from an updated version of the table in Lange et al, Nature Reviews Cancer 2011 (https://www.mdanderson.org/documents/Labs/Wood-Laboratory/human-dna-repair-genes.html) [116]. These are annotated with the pathways they are involved with (eg. nucleotide-excision repair, mismatch repair). I defined 'rare' variant as those with an allele frequency of <0.001 for heterozygous variants and those with an allele frequency of <0.01 for homozygous variants in both 1000 Genomes as well as across the 100kGP cohort.

3.2.9 Estimating the fraction of variance explained

To estimate the fraction of germline mutation variance explained by several factors, I fit the following Poisson GLMs with an identity link. I would expect data quality to correlate with the number of DNMs detected so to reduce this variation I used a subset of the 100kGP dataset which had been filtered on some base quality control (QC) metrics by the Bioinformatics team at GEL:

- cross-contamination $< 5\%$
- mapping rate $> 75\%$
- mean sample coverage > 20
- insert size <250

I then included the following variables to try and capture as much of the residual measurement error which may also be impacting DNM calling. In brackets I have given the corresponding variable names used in the models below:

- Mean coverage for the child, mother and father ($child\_mean\_RD$, $mother\_mean\_RD$, $father\_mean\_RD$)
- Proportion of aligned reads for the child, mother and father ($child\_prop\_aligned$, $mother\_prop\_aligned$, $father\_prop\_aligned$)
- Number of SNVs called for child, mother and father ($child\_snvs$, $mother\_snvs$, $father\_snvs$)
- Median VAF of DNMs called in child ($median\_VAF$)
- Median 'Bayes Factor' as outputted by Platypus for DNMs called in the child. This is a metric of DNM quality ($median\_BF$).

The first model I fit only included parental age:

$$E(Y) = \beta_0 + \beta_1 paternal\_age + \beta_2 maternal\_age$$

The second model also included data quality variables as described above:

$$\begin{aligned} E(Y) =& \beta_0 + \beta_1 paternal\_age + \beta_2 maternal\_age + \\ & \beta_3 child\_mean\_RD + \beta_4 mother\_mean\_RD + \beta_5 father\_mean\_RD + \\ & \beta_6 child\_prop\_aligned + \beta_7 mother\_prop\_aligned + \beta_8 father\_prop\_aligned + \\ & \beta_9 child\_snvs + \beta_{10} mother\_snvs + \beta_{11} father\_snvs + \\ & \beta_{12} median\_VAF + \beta_{13} median\_BF \end{aligned}$$

The third model included a variable for excess mutations in the 14 confirmed hypermutated individuals (hm_excess) in the 100kGP dataset. This variable was the total number of mutations subtracted by the median number of DNMs in the cohort (65), $Y_{hypermutated} - median(Y)$ for these 14 individuals and 0 for all other individuals.

$$\begin{aligned} E(Y) =& \beta_0 + \beta_1 paternal\_age + \beta_2 maternal\_age + \\ & \beta_3 child\_mean\_RD + \beta_4 mother\_mean\_RD + \beta_5 father\_mean\_RD + \\ & \beta_6 child\_prop\_aligned + \beta_7 mother\_prop\_aligned + \beta_8 father\_prop\_aligned + \\ & \beta_9 child\_snvs + \beta_{10} mother\_snvs + \beta_{11} father\_snvs + \\ & \beta_{12} median\_VAF + \beta_{13} median\_BF + \beta_{14} hm\_excess \end{aligned}$$

The fraction of variance (F) explained after accounting for Poisson variance in the mutation rate was calculated in a similar way to Kong et al using the following formula[111].

$$F = \frac{pseudo\text{-}R^2}{1 - \frac{\bar{Y}}{Var(Y)}}$$

I used McFadden's *pseudo-*R^2 as I was fitting a Poisson GLM. I also repeated these analyses fitting an ordinary least squares regression, as was done in Kong et al, using the R^2 from that and got comparable results. To calculate a 95% confidence interval I used a bootstrapping approach. I sampled with replacement 10,000 times and extracted the 2.5% and 97.5% percentiles.

Simulations to explore effect of non-random paternal age sampling

To look at the effect that non-random paternal age sampling has on the fraction of germline mutation rate explained I performed the following simulation:

I first simulated a random sample as follows 5,000 times:

- Randomly sample 78 trios
- Fit OLS of $E(Y) = \beta_0 + \beta_1 paternal\_age$
- Estimated fraction of variance (F) as described above

I then simulated a random sample as follows 5,000 times:

- Sample 78 trios as follows:
  - Sample $\frac{3}{4}$ of the 78 trios from the set of trios were paternal age falls into the top of bottom quartile (paternal age <29 or $\geq$37 years)
  - Sample $\frac{1}{4}$ of the 78 trios from those in the two middle quartiles (29$\leq$ paternal age < 37 years)
- Fit OLS of $E(Y) = \beta_0 + \beta_1 paternal\_age$
- Estimated fraction of variance (F) as described above

3.2.10 Analysis of contribution of rare variants in DNA repair genes

I fit 8 separate regressions to assess the contribution of rare variants in DNA repair genes. These were across three different sets of genes: variants in all DNA repair genes, variants in a subset DNA repair genes known to be associated with BER, MMR, NER or a DNA polymerase and variants within this subset that have also been associated with a cancer phenotype. For this I downloaded all ClinVar entries as of October 2019 and searched for germline 'pathogenic' or 'likely pathogenic' variants annotated with cancer [115]. I tested both nonsynonymous and PTVs for each set.

To assess the contribution of each of these sets I created two binary variables per set indicating a presence or absence of a maternal or paternal variant for each individual and then ran a poisson regression for each subset including these as independent variables along with hypermutation status, parental age and QC metrics as described in the previous section.

3.3 Results

3.3.1 Examining the effect of PTVs in *MBD4* on germline mutation rate

To investigate genetic variants that may impact germline mutation rate I at first took a genotype-driven approach. I examined the effect of PTVs in the known cancer mutator gene *MBD4* which are associated with a three-fold elevated CpG>TpG mutation rate in tumours. The CpG signature should be seen in both maternally and paternally derived mutations however I would expect to have more power to detect this elevated mutation rate in paternal germlines due to the larger number of paternal mutations. To this end, I identified 13 paternal carriers of *MBD4* PTVs within the DDD study that had a sufficient amount of remaining sample for sequencing and whole genome sequenced them and their parents. DNMs were called and filtered as described in the methods and post-filtering the individuals had an average of 67 DNMs per person. This is not elevated compared to what we would expect under the null and no individual had a significantly large number of DNMs. The mutational spectra looked normal for every individual (Figure 3.1c) and, using the proportion of CpG>TpG mutation expected from previous studies in healthy trios[173], there was no significant increase in the number of CpG>TpG mutations (p = 0.56, χ^2-test, Figure 3.1a). The 95% confidence interval around the CpG mutation rate 'multiplier' is 0.90 to 1.22 (ratio of two proportions), so I can confidently exclude that there is more than a 22% increase in the CpG mutation rate. This demonstrates that *MBD4* PTVs are unlikely to have a similar effect in the germline as in the soma.

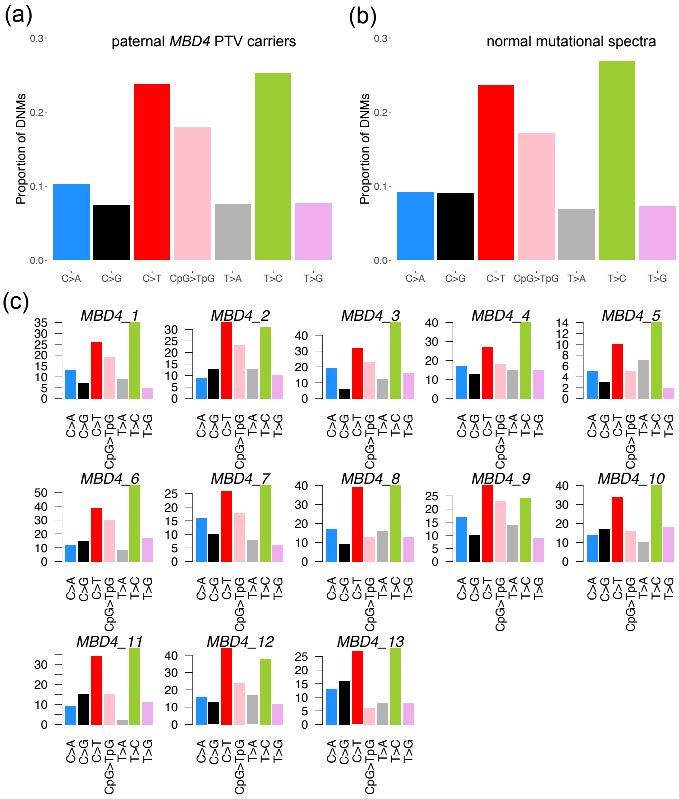


Fig. 3.1 Comparing the mutational Spectra of DNMs across the 13 paternal *MBD4* paternal PTV carriers (a) with the expected proportion of mutations (b) in each mutation type taken from Rahbari et al. [173] (c) The invididual mutational spectra demonstrating that no one individual has an elevated number of CpG>TpG mutations

3.3.2 Identifying germline hypermutators

For the phenotype driven approach I aimed to identify germline hypermutators. For this, I sought to identify offspring with an unusually large number of DNMs in exome-sequenced parent offspring trios in the DDD study and subsequently whole-genome sequenced trios in the rare disease cohort of the 100kGP. For the 100kGP data, this began with extensive DNM filtering that allowed me to explore additional properties of germline mutation variation including parental age. This was an important factor to account for in my downstream analyses. I was also able to explore differences in mutational spectra for maternally versus paternally derived DNMs.

Properties of *de novo* mutations in the 100kGP dataset

DNMs were called in 13,949 parent-offspring trios, across 12,609 families, as part of the rare disease cohort in the 100kGP dataset. After extensive filtering in collaboration with the bioinformatics team at Genomics England Limited (GEL), this resulted in a total of 999,939 DNMs: 921,433 *de novo* SNVs (dnSNVs) and 78,506 *de novo* indels (dnIndels). IGV examination of 300 random SNVs and 250 random indels demonstrated 95% true positive rate for SNVs and 90% true positive rate for indels. The VAF distribution and mutational spectra of these mutations are as expected (Figure 3.2). The median number of DNMs per individual was 65 for SNVs and 5 for indels, the number of SNVs and indels looked normal apart from some extreme outliers (Figure 3.3).

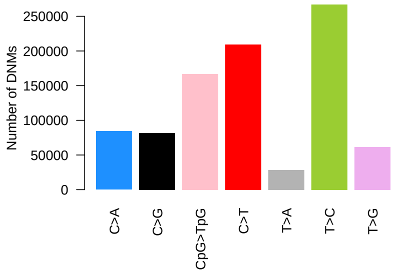


Fig. 3.2 Mutational Spectra of all DNMS called in the 100kGP cohort

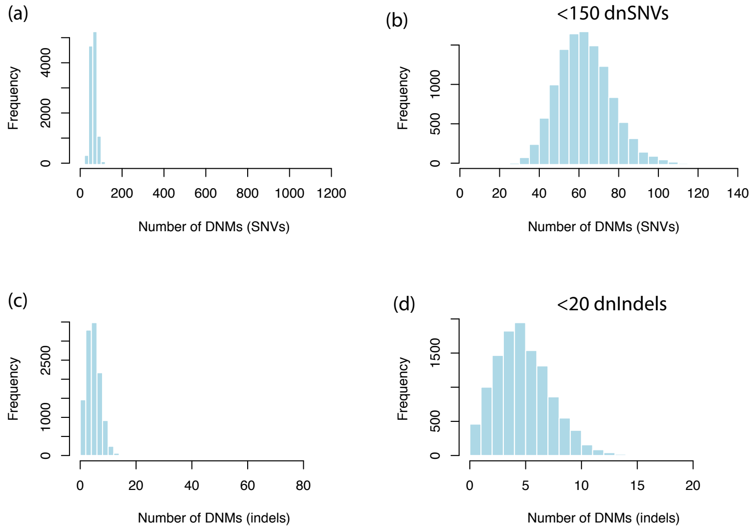


Fig. 3.3 Distribution of number of *de novo* SNVs for all individuals (a) and those with <150 DNMs (b). Distribution of number of *de novo* InDels per person for all individuals (c) and those with <20 indels (d)

Effect of parental age on germline mutation rate and parental differences in mutational spectra

To assess the effect of parental age on the germline mutation rate I ran a Poisson regression of the number of DNMs in the offspring on both maternal and paternal age at birth. This was done separately for SNVs and indels. Both paternal and maternal age were significantly associated with the number of *de novo* SNVs, I found an increase of 1.27 dnSNVs/year of paternal age (CI: 1.24-1.39, p $<10^{-300}$) and an increase of 0.35dnSNVs/year of maternal age (CI: 0.32-0.39, p $= 2.8 \times 10^{-80}$) (Figure 3.4a). These estimates agree with previous results reported in the literature ([234, 98]).

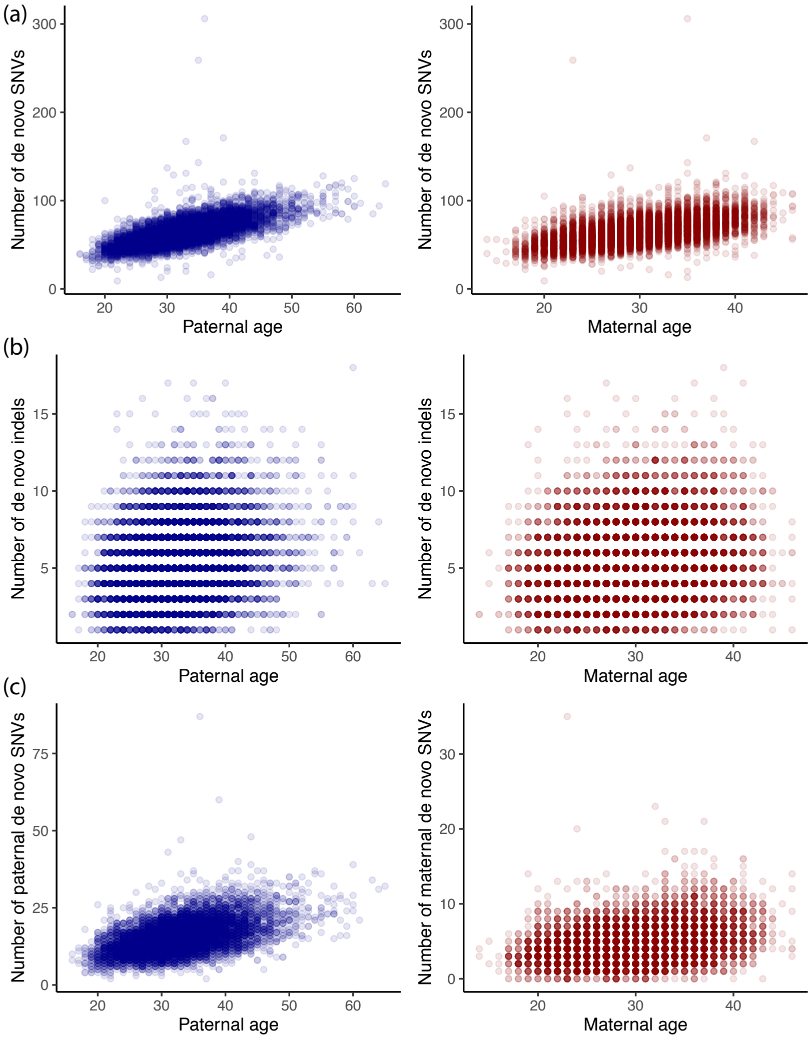


Fig. 3.4 Paternal and maternal age against the number of (a) dnSNVs, (b)dnInDels. (c) Paternal age against number of paternally phased dnSNVs and maternal age against number of maternally phased dnSNVs

I was able to phase 225,854 dnSNVs and the ratio of paternal to maternal DNMs was 3.29 across the dataset, 77% of phased DNMs were paternal in origin which agrees with previous studies [54, 173, 62]. I regressed the number of paternal mutations on paternal age and similarly the number of maternal mutations on maternal age. The effect estimates were not significantly different to the unphased results: 1.24 paternal dnSNVs/year of paternal age (CI:1.20-1.28, p $<10^{-300}$ and 0.38 maternal dnSNVs/year of maternal age (CI: 0.35-0.40, p $= 1.6 \times 10^{-211}$)(Figure 3.4c). Paternal and maternal age were also significantly associated with the number of dnIndels. I found that there was an increase of 0.078 dnIndels/year of paternal age (CI:0.068-0.087, p=1.96×10^{-64}) and a smaller increase of 0.021 dnIndels/year of maternal age (CI: 0.010-0.0031 p = 1.2×10^{-4})(Figure 3.4b). The ratio of paternal to maternal mutation increases for SNVs and indels were very similar, 3.7 for SNVs and 3.6 for indels.

Using the set of phased mutations I was also able to examine differences in properties between paternally and maternally derived DNMs. I found that the proportion of *de novo* mutations that phased paternally increased significantly with paternal age with a proportion increase of 0.0015 for every year of paternal age (p = 2.37×10^{-21}, Binomial regression) (Figure 3.5). This supports the idea that part of the paternal age effect is driven by replication errors as spermatogonial stem cells continue to divide after male puberty while female germ cells do not. However the effect size is small and the proportion of DNMs that phase paternally in the youngest fathers is ~0.75 and so replication errors alone do not fully explain the strong paternal bias.

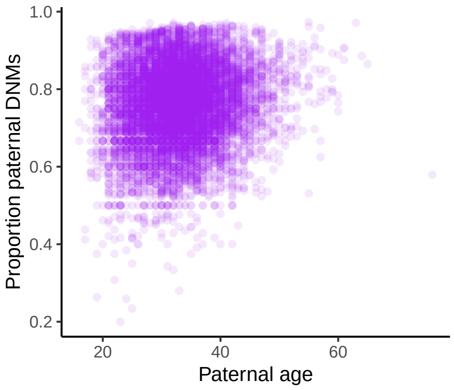


Fig. 3.5 Proportion of paternally phased DNMs against paternal age

I observed significant differences in the mutational spectra of paternally and maternally derived DNMs (Figure 3.6a). Maternally derived DNMs have a significantly higher proportion of C>T mutations while paternally derived DNMs have a significantly higher proportion of C>A,T>G and T>C mutations (p-values: 1.48×10^{-19},2.25×10^{-21},0.002, Binomial test).

These mostly agree with previous studies although the difference in T>C mutations was not previously significant [62]. To further understand the differences in the mutational profile, I extracted mutational signatures for maternally and paternally phased DNMs. These were then mapped on to known mutational signatures from COSMIC and found that the majority of the mutations could be explained by Signature 1 and 5 as has previously been observed in germline mutation (Figure 3.6b) [173]. I found that the proportion of mutations explained by Signature 1 was significantly greater in the paternal compared to maternal mutations although the difference was very slight (0.15 paternal vs 0.14 maternal, chi-sq test p = 4.53×10^{-6}).

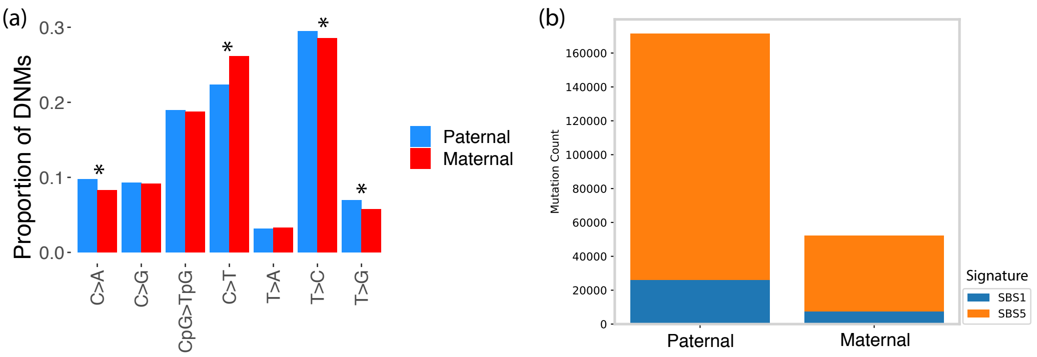


Fig. 3.6 (a) Mutational spectra for maternal vs paternal DNMs across 100kGP cohort. Significant differences (p<0.05/7) are marked with *. (b) Mutational signature decomposition for DNMs in maternally and paternally derived DNMs. Signatures extracted with SigProfiler. Colors correspond to COSMIC signatures.

Identifying hypermutated individuals in DDD and 100kGP

To identify hypermutated individuals in the DDD study, I analysed exome DNMs called in probands from 7,930 parent-offspring trios. This is a slightly larger set than the one described in Chapter 2 but the DNMs were called in the same way and subject to the same filters. To identify probands with an excess of exome DNMs it was important to account for parental age. I fit a Poisson generalized linear model (GLM) with maternal and paternal age as covariates and then looked for individuals that had both a high regression residual and a large absolute number of exome DNMs. After inspection of IGV plots, to ensure the exome DNMs appeared to be real, and ensuring that the child was related to both parents, I narrowed down the list to 10 trios. The 10 probands had 7-17 exome DNMs. It is important to note that not all DNMs detectable from WES fall within exons. The baits overlap with non-coding regions as the exome capture for DDD also had an additional 5MB of non-coding elements. These individuals were then whole genome sequenced to >30 mean depth using Illumina short-read sequencing. Due to a sample fail, I could only analyse 9 of the 10 trios.

ID	Number of SNVs	Number of InDels	Paternal age	Maternal age	SNV p-value	InDel p-value	transcriptional strand-bias	Phase P,M	Phase Ratio p-value	Hypermutation type
GEL_1	425	16	(30,35]	(20,25]	1.78E-68	9.18E-05	7.82E-23	129,1	4.19E-14	paternally_phased
GEL_2	368	6	(25,30]	(25,30]	2.45E-51	0.363	0.219	100,7	1.35E-06	paternally_phased
GEL_3	306	4	(35,40]	(30,35]	3.30E-30	0.745	0.078	87,5	3.86E-06	paternally_phased
DDD_1	277	6	25	37	NA	NA	3.29E-03	72,4	8.06E-07	paternally_phased
GEL_4	259	11	(30,35]	(20,25]	3.91E-21	0.028	0.608	37,35	1.00	post-zygotic
GEL_5	171	7	(35,40]	(35,40]	1.71E-06	0.381	0.096	58,4	2.85E-04	paternally_phased
GEL_6	167	7	(30,35]	(40,45]	1.06E-06	0.330	1	36,4	0.028	other
GEL_7	143	9	(30,35]	(30,35]	1.76E-04	0.129	0.039	23,17	0.998	post-zygotic
GEL_8	137	7	(25,30]	(25,30]	1.11E-04	0.274	0.141	33,11	0.680	other
GEL_9	131	6	(30,35]	(30,35]	9.10E-04	0.448	0.427	47,3	0.001	paternally_phased
GEL_10	131	13	(40,45]	(35,40]	6.35E-03	0.010	0.063	29,15	0.965	post-zygotic
GEL_11	131	9	(40,45]	(35,40]	8.95E-03	0.195	0.268	48,9	0.115	other
GEL_12	129	5	(30,35]	(25,30]	5.23E-04	0.547	0.091	43,0	1.11E-05	paternally_phased
GEL_13	114	3	(30,35]	(30,35]	9.91E-03	0.820	0.001	19,5	0.499	other
GEL_14	111	8	(25,30]	(25,30]	4.96E-03	0.155	2.54E-06	31,1	0.002	paternally_phased

Table 3.1 Properties of hypermutated individuals. Maternal and Paternal age is given in 5 year window for 100kGP as this information was not allowed to be extracted from the research environment due to privacy implications. However the regression was run on the exact ages and the parental age plots also share the exact ages. Phase column de notes the number of DNMs that were phased the paternally (P) and maternally (M).

DNMs were called from these trios using DeNovoGear[174] and were subject to a set of filters described in the Methods. One of these individuals was apparently hypermutated, with 277 DNMs, ~4 fold as many as expected, while the remaining individuals did not have remarkably high numbers of DNMs (median of 81 DNMs).

Identifying hypermutated individuals in 100kGP was more straight forward as the individuals had all been whole-genome sequenced from the outset. After regressing out paternal and maternal age on the number of dnSNVs, 27 individuals had residuals which were larger than the remaining residual distribution using a p-value threshold of 0.01. This threshold was used as opposed to the Bonferroni corrected threshold of 4×10^{-6} as I wanted to capture all possible hypermutated individuals. These individuals had 111-1379 apparent dnSNVs per person. These were extensively followed up to remove false positives. After careful examination of the distribution of these DNMs and their corresponding IGV plots I determined that 14 of these were truly hypermutated (Table 3.1). Here I focused on identifying hypermutated individuals with a large number of dnSNVs rather than dnIndels. This was because I had more confidence in the filtering of SNVs and it was easier to confirm that the supposed hypermutation was not due to a larger structural event that was miscalled. However I did also regress out parental age on the number of dnIndels per individual and calculate a corresponding p-value for whether the residuals were significantly larger than the rest of the cohort (InDel p-value in Table 3.1). Only one of the 14 was significant for indel hypermutation.

There were two main error modes for the 13 individuals that I determined were not truly hypermutated. For ten of these individuals it appears that a somatic deletion in the blood of one of the parents has occurred leading to a very high number of supposed DNMs being called in that region in the offspring. These individuals had some of the highest number of DNMs called (up to 1379 DNMs per individual). For each of these 10 individuals, the DNM calls all clustered to a specific region in a single chromosome. In this same corresponding region in the parent, I observed a loss of heterozygosity when calculating the heterozygous/homozygous ratio (Figure 3.7). In addition, many of these calls appeared

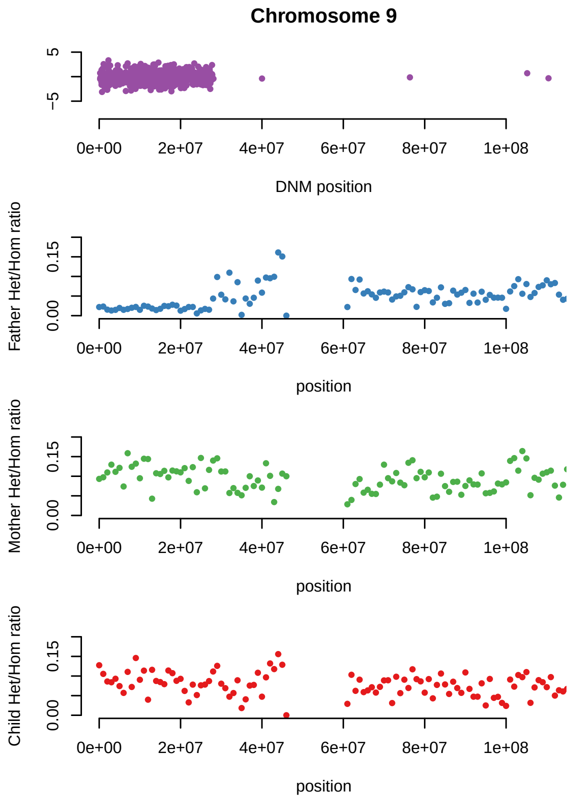


Fig. 3.7 Loss of transmitted allele example leading to false positive DNMs. Top plot shows the location of the called DNMs in the child on chromosome 9. The plots below show the heterozygous/homozygous ratio in the Father, Mother and Child showing a loss of heterozygosity in the father in the same region the DNMs have been called.

to be low level mosaic in that same parent. This type of event has previously been shown to create artifacts in CNV calls and is referred to as a ’Loss of Transmitted Allele’ event [175]. I removed two other individuals due to a high false positive rate of called DNMs upon examination of IGV plots and therefore these did not appear to be truly hypermutated. The last individual that I removed had 100 autosomal DNMs and the largest p-value very close to the threshold (p = 0.0099). The mutational spectrum was normal, no specific mutation type was significantly enriched, the VAF distribution was normal and the mutations did not phase more to one parent compared to what we would expect. This led me to believe that this may be an individual on the tail of the DNM count distribution rather than hypermutation.

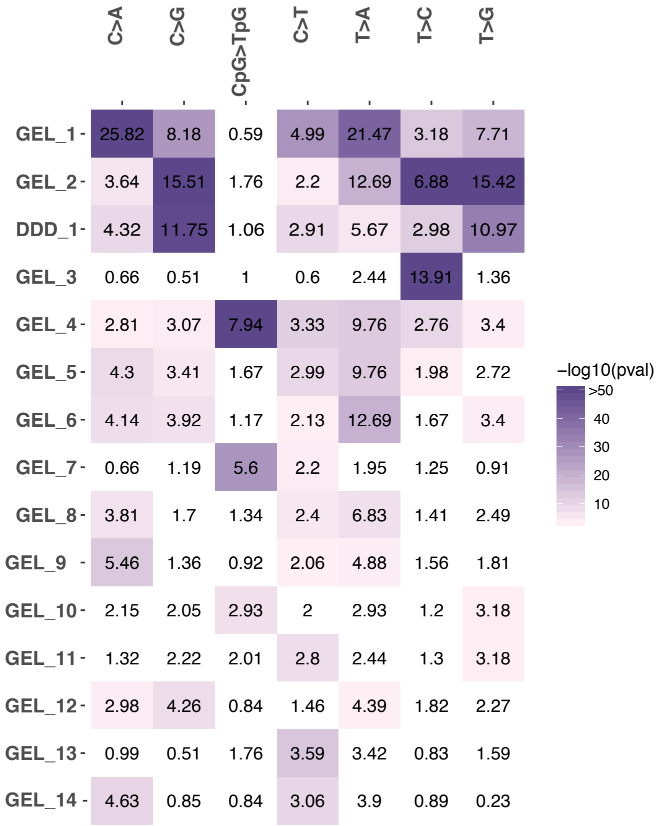
3.3.3 Characterising hypermutation in 15 individuals

The number of DNMs for each of these 15 hypermutated individuals across both DDD and 100kGP ranged from 111-425 which corresponds to a fold increase of 1.7-6.5 compared to the median number of DNMs per individual across the 100kGP cohort. For each of the 15 hypermutated individuals I explored various characteristics of their DNMs to uncover possible underlying causes of this mutator phenotype (Table 3.1). The mutational spectra varied widely (Figures 3.14,3.15) and I calculated the enrichment of each of these mutation types compared to the average number of mutations observed across the 100kGP cohort (Figure 3.8). I extracted mutational signatures for all of these individuals using SigProfiler (Figure 3.9a)[12]. I found that most of the DNMs mapped on to known mutational signatures in cancer (from COSMIC) however there was also a novel signature extracted (Figure 3.9b)[212]. In addition to mutational spectra, I analysed parental phase of the DNMs, transcriptional strand bias and VAF distributions. Upon examining these properties, I was able to categorise these individuals into three different groups.

Hypermutation due to parental hypermutator

The first of these groups comprised of individuals whose excess DNMs originated from a single parent. I was able to phase ~$\frac{1}{3}$ of DNMs in these individuals and found that for eight of the fifteen the DNMs phased to the father significantly more than what we would expect given the overall ratio of paternal:mutations across all individuals in the 100kGP cohort (p-<0.05/15, Binomial test, Table 3.1). An additional individual was nominally significant (GEL_6 p = 0.028). This implicates the father as a possible germline hypermutator. To try and identify possible genetic causes I searched for rare paternal variants in known DNA repair genes compiled from the literature. Defects in DNA repair are known to increase the mutation rate in the soma and therefore may have a similar effect in the germline. I found possible causal variants in two of these individuals (Table 3.2).

GEL_1 has the largest number of DNMs of all individuals, a ~7 fold enrichment compared to what we would expect. The mutational spectra demonstrates a high enrichment of C>A and T>A mutations (Figure 3.14a,3.8). From extracting mutational signatures I observed a large contribution from Signature 8 in COSMIC (Figure 3.9). This signature is associated with transcription-coupled nucleotide excision repair (TC-NER) and typically presents



	C>A	C>G	CpG>TpG	C>T	T>A	T>C	T>G
GEL_1	25.82	8.18	0.59	4.99	21.47	3.18	7.71
GEL_2	3.64	15.51	1.76	2.2	12.69	6.88	15.42
DDD_1	4.32	11.75	1.06	2.91	5.67	2.98	10.97
GEL_3	0.66	0.51	1	0.6	2.44	13.91	1.36
GEL_4	2.81	3.07	7.94	3.33	9.76	2.76	3.4
GEL_5	4.3	3.41	1.67	2.99	9.76	1.98	2.72
GEL_6	4.14	3.92	1.17	2.13	12.69	1.67	3.4
GEL_7	0.66	1.19	5.6	2.2	1.95	1.25	0.91
GEL_8	3.81	1.7	1.34	2.4	6.83	1.41	2.49
GEL_9	5.46	1.36	0.92	2.06	4.88	1.56	1.81
GEL_10	2.15	2.05	2.93	2	2.93	1.2	3.18
GEL_11	1.32	2.22	2.01	2.8	2.44	1.3	3.18
GEL_12	2.98	4.26	0.84	1.46	4.39	1.82	2.27
GEL_13	0.99	0.51	1.76	3.59	3.42	0.83	1.59
GEL_14	4.63	0.85	0.84	3.06	3.9	0.89	0.23

Fig. 3.8 Enrichment (observed/expected) of mutation type for hypermutated individuals. Sample names on the y axis, mutation type on the x axis. The enrichment is colored by the -log10(enrichment p-value) which was calculated using a poisson test comparing the average number of mutations in each type across all individuals in the 100kGP cohort. White coloring indicates no statistically significant enrichment (p-value <0.05/(15 × 7))

with transcriptional strand bias on the untranscribed strand. This agrees with the strong transcriptional strand bias I observed in GEL_1 ($p = 7.8 \times 10^{-23}$, Poisson test, Figure 3.10). This individual was also the only hypermutated individual that also had a significantly increased number of *de novo* indels ($p = 9.18 \times 10^{-5}$, Table 3.1). In my analysis of rare paternal variants in DNA repair genes, I identified a homozgyous stop gained variant in the gene *XPC* (Table 3.2). *XPC* is involved in the early stages of the nucleotide-excision repair (NER) pathway. NER is the main pathway for removing various types of DNA lesions such as those induced by UV light as well as other chemical adducts. There are several rare autosomal recessive syndromes that are a result of defects in NER; these include Cockayne syndrome, trichothiodystrophy and xeroderma pigmentosum [29]. The paternal variant that I identified is annotated as pathogenic for xeroderma pigmentosum in ClinVar and there are no observed homozygotes in the genome aggregation database (gnomAD AF

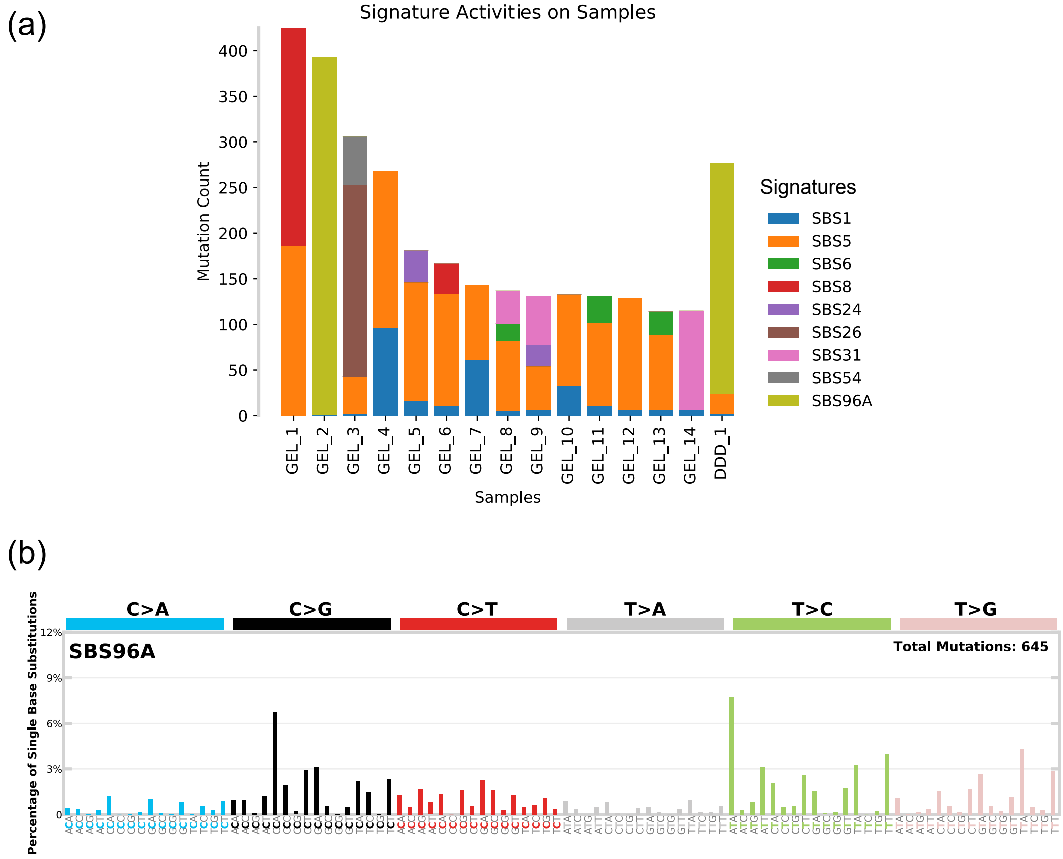


Fig. 3.9 Mutational signature decomposition for DNMs in hypermutated individuals. (a) Signatures extracted with SigProfiler. Colored by signatures number, these numbers correspond to COSMIC mutational signatures apart from SBS96A with is a novel signature. (b) The novel signature extracted which contributes heavily to GEL_2 and DDD_1.

= 2.2×10^{-5})[115, 102]. Upon contact with the corresponding clinician for this patient it was confirmed that the father has been diagnosed with the disorder. Patients with xeroderma pigmentosum have a high risk of developing skin cancer due to their impaired ability to repair UV damage and are also known to be at a higher risk of developing other cancers [123, 169]. *XPC* deficiency has been associated with a similar mutational spectrum to the one we observe in GEL_1. A recent study observed increased Signature 8 mutations in a human intestinal organoid culture in which *XPC* was deleted using CRISPR-Cas9 gene-editing, although transcriptional strand bias was not observed here[92]. The same study observed that genomes of NER-deficient breast tumors show an increased contribution of Signature 8 mutations compared with NER-proficient tumors. There is little previous evidence of the effect of *XPC* deficiency on germline mutation in humans, although a previous study has

ID of child	Chrom	Position (hg38)	Ref	Alt	Csq	Paternal genotype	Gene	DNA repair pathway	Gnomad AF	Pathogenicity evidence
GEL_1	3	14165549	G	A	stop_gained	1/1	*XPC*	NER	2.2e-5	Pathogenic for xeroderma pigmentosum in ClinVar
GEL_5	16	83139	G	A	missense	1/1	*MPG*	BER	9.57e-5	CADD score 27.9; likely interacts with DNA

Table 3.2 Possible paternal mutator variants

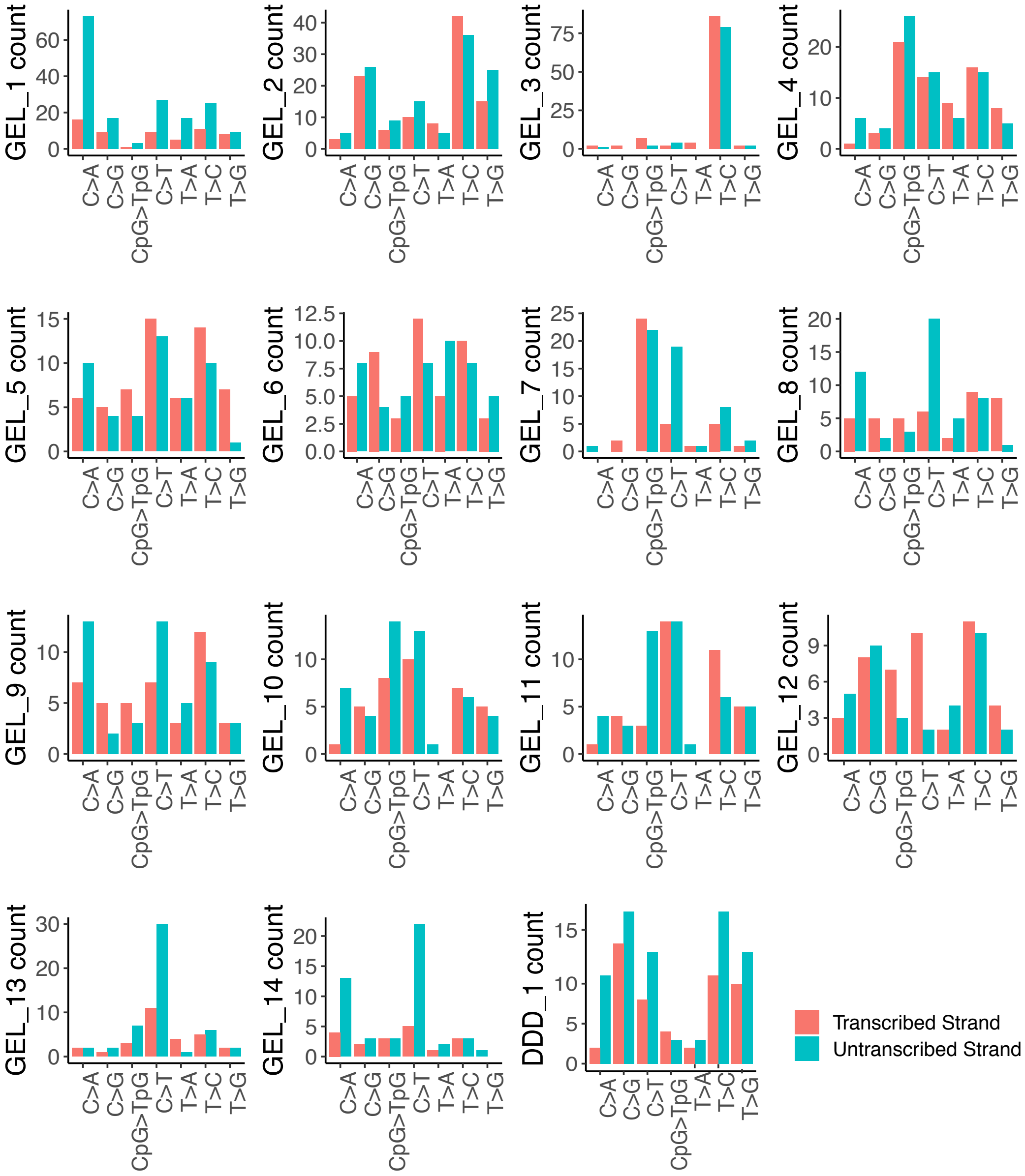


Fig. 3.10 Transcriptional strand bias for DNMs in hypermutated individuals

shown that Xpc deficient (-/-) male mice have a significantly increased germline mutation rate at two STR loci compared to heterozygous *XPC* (+/-) and wild-type (+/+) mice which may indicate a mutator phenotype [145].

GEL_3 has a ~5 fold enrichment of the number of DNMs. These DNMs exhibit a very distinct mutational spectrum with a ~14 fold increase in C>T mutations but no significant enrichment for any other mutation type (Figure3.14d, Figure 3.8). Extraction of mutational signatures revealed that the majority of mutations mapped onto Signature 26 from COSMIC (Figure 3.9a). This signature is associated with defective mismatch repair. In my analysis of paternal variants, I identified a rare homozygous missense mutation in the gene *MPG* (Table 3.2). *MPG* encodes for a DNA glycosylase which is involved in the recognition of base lesions, including alkylated and deaminated purines, and initiation of the base-excision repair (BER) pathway. The paternal variant I identified has an allele frequency of 9.8×10^{-5} in gnomAD with 0 observed homozygotes. The Combined Annotation Dependent Depletion (CADD) score, for this variant is 27.9 and the amino acid residue is highly conserved (conservation = 1 from 172 aligned protein seqs from VarSite) [117]. An analysis of its position in the context of the protein by James Stephenson, a post-doc in the group, revealed it forms part of the substrate binding pocket and is likely interacting with DNA (Figure 3.11) [118]. Studies in yeast have demonstrated that overexpression of *MPG* can lead to a mutator phenotype and that variants that alter other amino acids in the substrate binding pocket, and alter substrate specificity, can result in an increase in the mutation rate of either point mutations or STRs[61, 48]. Another study found that *Mpg*(-/-) mice treated with methyl methanesulfonate resulted in >3 times *hprt* mutations in splenic T lymphocytes compared to wildtype also demonstrating that there can be a mutagenic effect [47].

GEL_2 and DDD_1 have a similar number of DNMs which are significantly more paternal in origin than expected (Table 3.1). The mutational spectra of the DNMs in these individuals are very similar and the cosine similarity between their spectra is 0.79 (Figure 3.14). In my analysis of mutational signatures, a novel signature was extracted which these two individuals share. This does not map onto any known signatures in COSMIC and is characterised by an enrichment of C>G and T>G mutations (Figure 3.9a,b). In my analysis of paternal variants in DNA repair genes I found that the father of DDD_1 has a rare heterozygous missense variant in *BRCA2* and a heterozygous stop gained mutation in the gene *NTHL1*. The *BRCA2* variant has an allele frequency of 0 in gnomAD and is annotated as a variant of uncertain significance (VUS) for breast cancer in ClinVar. It has conflicting interpretations of pathogenicity from different tools (SIFT:'Tolerated', PolyPhen-2:'Probably Damaging'). *BRCA2* is involved in the homologous recombination repair pathway that mends double strand breaks and so defects in *BRCA2* in cancer are known to lead to an increase in

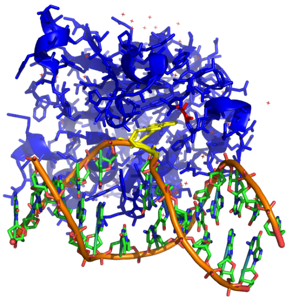


Fig. 3.11 Position of *MPG* missense variant (residue in red) in GEL_3 in the context of the protein (blue). Residue forms part of the binding pocket and image demonstrates its proximity to DNA (orange). Image courtesy of James Stephenson

the number of indel mutations as well as SNVs. DDD_1 does not map on to any mutational signatures associated with defects in *BRCA2* and does not have a significantly increased number of dnIndels and so this variant does not look convincingly causal [159]. *NTHL1* is a gene involved in the BER pathway and germline homozygous mutations in this gene have been associated with multiple cancers [231]. The paternal variant in *NTHL1* has an allele frequency of 1.42×10^{-3} in gnomAD and there were an additional 23 fathers in the DDD study that had this same variant. I examined the mutational spectra across all the DNMs from their offspring and found they were normal and did not have this distinctive signature so this is unlikely to be the sole cause of hypermutation. There were no putative damaging paternal variants in DNA repair genes for GEL_2. Since these two individuals shared this mutational signature I looked for an intersection of genes in which both individuals had rare nonsynonymous paternal variants. For this I looked across all genes, not restricted to DNA repair genes, but found no overlap. In the corresponding clinician's additional notes for patient DDD_1 it has been noted that the father has undergone treatment for Hodgkin's Lymphoma twice. This may be a result of a paternal mutator variant also having an effect in the paternal soma and increasing cancer risk or the hypermutation in the child could be due to damage incurred in the father's germline during cancer treatment. The mutational signature does not resemble known signatures associated with Hodgkin's Lymphoma in cancer or known chemotherapeutic signatures (Table 3.1) [166, 164]. The father does not have any known germline variants that are associated with elevated risk of Hodgkin's Lymphoma although there are other germline *BRCA2* variants that can increase risk[122]. I am currently following up with the corresponding clinician to confirm these cancer treatments occurred prior to the conception of the child and what these treatments were. I am also following up with GEL_2 to see if their father has had cancer or undergone treatment as well.

For the remaining five individuals that may have a paternal hypermutator, I was not able to identify any putatively causal paternal variants. The mutational signatures in these individuals have various compositions which may indicate the mechanisms in which the DNMs arose. For example the DNMs in GEL_14 map mostly onto Signature 31 which is associated with transcription coupled NER (Figure 3.9). The significant transcriptional strand bias ($p = 2.54 \times 10^{-6}$) in the DNMs would support this mechanism however I did not observe any nonsynonymous rare variants in genes known to be involved in NER. For these five individuals, a paternal mutator variant may fall into a gene not currently associated with DNA repair or may be non-coding. I searched for rare recessive paternal variants in all genes across these five individuals but there was nothing immediately notable. Other explanations may be that the variant may be germline specific and so not detectable in blood, the hypermutation may be due to an environmental mutagen that has impacted the paternal germline or there may be a gene by environment interaction that results in increased mutation rate.

Post-zygotic hypermutation

The second group of hypermutated individuals consists of those where the hypermutation appears to have occurred post-zygotically. I examined the distribution of the VAF in the DNMs for each individual. I found that for three of these individuals (GEL_4, GEL_7 and GEL_10) the VAF distribution was not centered around 0.5 (Figure 3.12). The proportion of DNMs with VAF<0.4 was significantly higher than compared to the distribution of all DNMs across all individuals in GEL_4 ($p = 1.5 \times 10^{-51}$, Binomial test) and GEL_10 ($p = 2.4 \times 10^{-4}$) and nominally significant in GEL_7 ($p = 0.02$). For all three of these individuals, the mutations phased evenly between the maternal and paternal chromosome. This indicates that these mutations most likely occurred post-zygotically and are less likely to be due to a parental hypermutator. All three of these individuals are most strongly enriched for CpG>TpG mutations and have a large contribution of mutations from Signature 1 in COSMIC (Figure 3.9, Figure 3.8).

Other sources of hypermutation

The third group of hypermutated individuals included the remaining 4 hypermutated individuals. The DNMs in these individuals did not phase overwhelmingly to a single parent and the VAF distributions did not indicate a large number of post-zygotic mutations (Figure

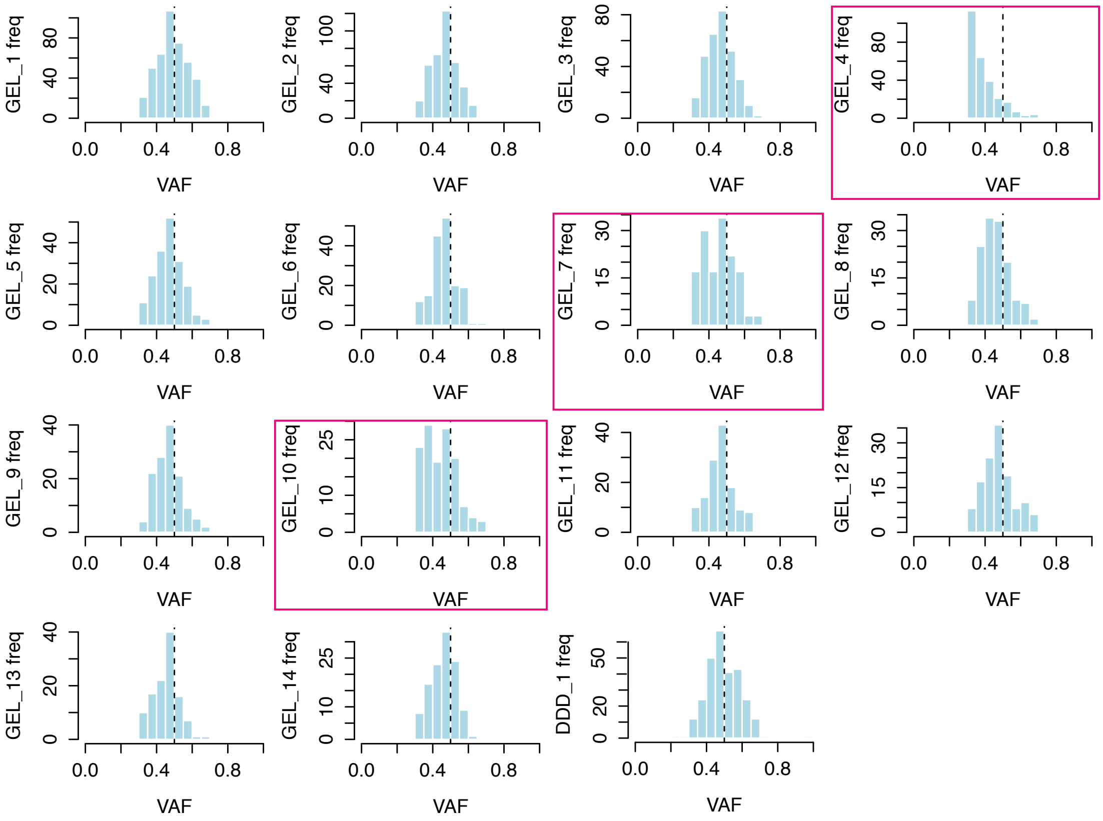


Fig. 3.12 Distribution of variant allele fraction (VAF) for DNMs in hypermutated individuals. The vertical line indicates 0.5 VAF. The three plots highlighted in pink are those where the DNMs appear post-zygotic.

3.12, Table 3.1). They did appear to have mutational spectra that are in different proportions to what we would expect. I observed different levels of enrichment across mutation types compared to expected (the average number of mutations across 100kGP) (Figure 3.8) however these were not as striking. The observed elevated germline mutation rates may be due to a combination of polygenic effects in the parents, shared mutagenic environment for the parents or an interaction between the two.

3.3.4 Fraction of germline mutation rate variation explained

Work from Kong et al. studying 78 trios previously estimated that paternal age accounts for >95% of the variation surrounding germline mutation rate after accounting for Poisson variation [111]. Using a similar approach I fit several GLMs including variables for parental age and hypermutation status and calculated the fraction of variance explained in the 100kGP dataset. To mitigate the effect of data quality this analysis was performed on a subset of

7,700 trios that had been filtered on basic QC metric such as coverage and mapping rate. I also removed the false positive hypermutated individuals that I identified. The details of this can be found in the Methods. I first fit a model that only accounts for parental age and found this explained 70% of the variation of the number of mutations per individual.

This estimate of 70% is considerably lower than the previous estimate from Kong et al and there may be several explanations for this. Firstly, due to the much larger size of the dataset, I was unable to verify the DNMs to the same degree as in the Kong et al paper which was performed on 78 parent offspring trios. I estimated the true positive rate of the called DNMs to be 0.95, therefore the variance may be overestimated. This analysis was done on a subset of higher quality samples to mitigate this but to account for additional measurement error, which may correlate with the number of DNMs called, I also included coverage, mapping and variant calling metrics in my regression models and found this explained ~3% of variation. Secondly, Kong et al. may be slightly underestimating germline mutation rate variation due to the fact that the 78 trios in the paper also included multi-sibling families which we may expect to have more similar number of DNMs than unrelated trios, this would inflate the variation explained. Thirdly, if the trios selected for the Kong et al. analysis were selected non-randomly with respect to paternal age then this could conflate the variance explained by that variable. I performed simulations in the 100kGP dataset where I sampled trios either randomly across the population or more heavily towards the tails (disproportionate amount of young/old fathers) and found that heavier tail sampling significantly increased the median proportion of variation explained from 0.78 to 0.82 ($p = 5.7 \times 10^{-61}$, Wilcox test). While this may contribute to the discrepancy, it is unlikely to fully explain the much higher fraction of variance explained by Kong et al. Finally, by repeated random sampling of 78 trios from the much larger 100kGP data I observed that in such a small dataset estimates of the variance explained varies considerably by chance, and that although the median estimate of variance explained was 0.78, I observed an estimate of variance explained similar or greater to that observed by Kong et al in 7% of simulations. This suggests that Kong et al could have over-estimated the true variance explained by parent age by chance, and that the uncertainty in their estimate was much greater than they estimated.

In addition to parental age and data quality I also included in the regression a variable accounting for the excess number of mutations in individuals I have identified and confirmed as being hypermutated. I found that this accounted for an additional 8% of variation. In total, this means that 20% (17%-22%, Bootstrap 95% confidence interval) of variation remains unaccounted for of which there may be several contributors. Variants in genes involved in DNA repair are implicated here as possible causes of hypermutation therefore they may also play a role in the remaining germline mutation rate variation. In addition, polygenic effects, environmental mutagens and gene by environment interactions may also contribute.

Impact of variants in DNA repair genes across cohort

To assess whether rare variants in genes known to be involved in DNA repair pathways impact germline mutation rate more generally, I looked across the whole 100kGP cohort. I curated three sets of variants that have increasing likelihoods of impacting germline mutation rate. For all three sets I considered both all nonsynonymous variants and restricting these to just PTVs. The first set was the least stringent set including 186 known DNA repair genes which is the same set described earlier. For this set I also separately considered the impact of rare homozgyous variants in these genes (the counts were too small to assess in the subsequent groups). The second set was restricted to DNA repair genes encoding components of the DNA repair pathways most likely to create SNVs. For this I chose the 66 genes that were known to be associated with the BER, NER and mismatch repair (MMR) pathways as well as DNA polymerases. Again, I looked at the impact of both nonsynonymous and just PTVs on germline mutation rate. The third set were variants within this second set that have also been associated with an increased risk of cancer. This was created by considering variants that are annotated as ’pathogenic’ or ’likely pathogenic’ germline variants for any cancer phenotype in ClinVar. I found that for all eight regressions that I ran there was no statistically significant effect after Bonferroni correction (Table 3.3, Figure 3.13). The only effect that

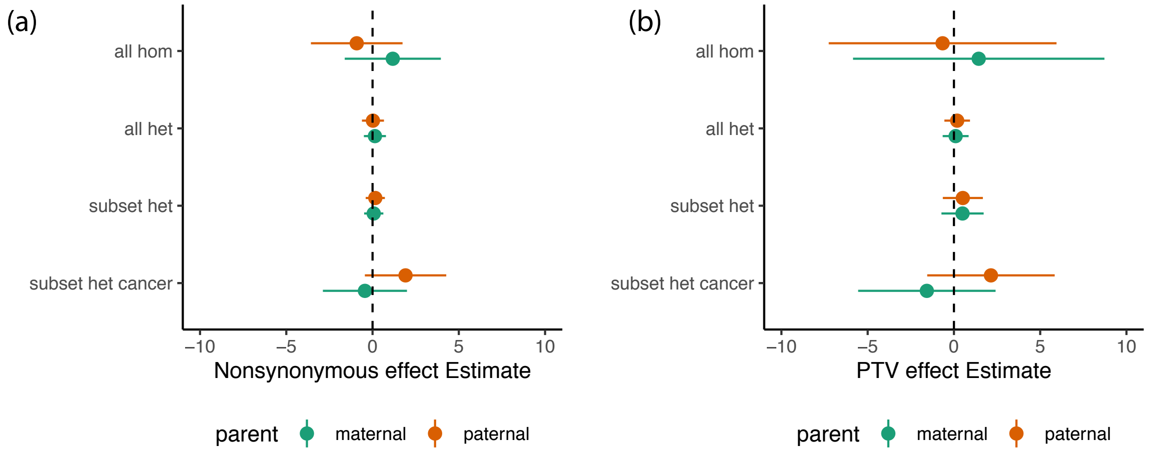


Fig. 3.13 Impact of rare variants in DNA repair genes on germline mutation rate. Poisson regression effect estimates for binary variables of having a parental variant in genes known to be involved in DNA repair. (a) considered all nonsynonymous variants in the subsets (b) is restricted to PTVs.

was nominally significant was for paternal nonsynonymous variants known to be associated with cancer phenotypes (p = 0.018) and this only explained an additional 0.03% of variance. This demonstrates that rare variants in DNA repair genes do not explain a large amount of the remaining variation in germline mutation rate. To detect more subtle effects of these variants other analytical approaches will need to be explored. The role of genetic variation, not restricted to these genes also needs to be investigated.

Variant subset	Consequence	Genotype	Paternal count	Paternal Effect	Paternal p-value	Maternal count	Maternal Effect	Maternal p-value
all DNA repair	nonsynonymous	het	5865	0.023	0.915	5916	0.137	0.526
	PTV	het	1203	0.187	0.456	1153	0.099	0.697
	nonsynonymous	hom	78	-0.917	0.307	71	1.174	0.213
	PTV	hom	13	-0.657	0.769	11	1.437	0.560
subset DNA repair	nonsynonymous	het	3076	0.159	0.398	2928	0.069	0.715
	PTV	het	434	0.516	0.189	391	0.498	0.229
germline cancer	nonsynonymous	het	103	1.912	0.017	97	-0.442	0.592
	PTV	het	41	2.145	0.086	35	-1.570	0.244

Table 3.3 Impact of parental rare variants in DNA repair genes on germline mutation rate

3.4 Discussion

Germline hypermutation is an uncommon but important phenomenon which can impact the health of subsequent generations. In this chapter, I identified 15 individuals from ~20,000 parent-offspring sequenced trios in the DDD study and 100kGP with a significant 2-7 fold increased number of DNMs compared to expected. For 3 of these individuals the excess mutations appear to have occured post-zygotically however for the majority (8) of these hypermutated individuals, the excess DNMs phased paternally implicating the father as a potential germline hypermutator. I identified possible paternal mutator variants in two of these individuals. These were rare nonsynonymous homozygous variants in two genes known to be involved in DNA repair, *XPC* and *MPG*. The missense variant in *MPG* is likely damaging however functional follow up is necessary here to assess whether and how it may disrupt the BER pathway and create such a distinctive mutational spectrum. A collaborator is currently carrying out functional assays to interrogate the impact of the change in this residue. The father carrying the *XPC* PTV has been diagnosed with xeroderma pigmentosum (XP) which carries a very high risk of skin cancer as well as an increased risk of other cancers.

It is well established that defects in DNA repair genes can increase the somatic mutation rate and elevate cancer risk [105]. The findings in this chapter imply that the germline can be similarly affected and that defects in DNA repair can lead to a dramatic increase in germline mutation rate. However defects in DNA repair pathways do not always appear to behave similarly in the soma and the germline. I interrogated protein-truncating variants in an established cancer mutator gene, *MBD4*, and found they did not have a detectable effect in the germline [232]. I also looked at the impact of parental nonsynonymous variants in DNA repair genes on the number of DNMs in offspring across the 100kGP cohort and did not find a significant difference. Paternal variants that have previously been associated with a cancer phenotype were nominally significant but having one of these variant only amounted to an estimated increase of approximately ~2 DNMs in the child. If only a subset of these variants have an impact in the germline this would dilute our power to detect an effect and it is likely we will need both larger sample sizes as well as a more stringently curated set of variants to investigate this further. There are also likely to be pathways that impact the germline more than the soma and uncovering the genes and associated variants in these genes will be more challenging.

A limitation to the approach I took in this chapter is that I used DNMs of a single offspring as a proxy for the germline mutation rate of both parents. Aside from sequencing large families, directly sampling the germline would be more reliable in estimating individual mutation rate. Sequencing oocytes is difficult to do at a large scale due to the invasive and costly procedure needed to sample only a few eggs. Moreover, I did not observe a significant maternal bias in any of the hypermutated individuals. Since the mother contributes only a quarter of a child's DNMs on average, I may be less powered to detect an increase in maternal DNMs. The maternal germline may also be more protected to mutator variants as oocytes stop replicating during gestation while spermatagonial stem cells continue to replicate throughout a male's life and may be more vulnerable to impaired repair processes due to uncorrected replication errors. Sperm is more feasible to sample at scale and would be an important resource to estimate individual male mutation rate variation. At a smaller scale we are currently following up with Genomics England Limited in order to recontact the likely paternal hypermutators to collect sperm for single-cell sequencing. This will allow us to interrogate whether all sperm are affected equally by the hypermutation, the presence of mutator variants that are only present in the germline and improve our ability to extract mutational signatures on a larger number of mutations. Another useful next step would be to follow up more directly with parents with different DNA repair disorders, including those with pathogenic variants in *XPC*. Sequencing sperm or families of other male XP patients would allow us to see if germline hypermutation is observed in those with the same and other pathogenic variants in this gene. Variants in other other genes associated with XP (*XPA*, *XPB* etc.) might also be worth investigating. This information may be clinically useful for these patients as germline hypermutation means future children are at a higher risk of having a genetic disorder caused by a DNM.

It is important to note that to identify hypermutators I fit an ordinary linear regression of the number of DNMs on parental age and then applied a threshold on the studentized residuals to capture those with an unusually large number of DNMs. In part, this was for comparability to the Kong et al study, which used the same regression approach. The studentized residuals are expected to follow a t distribution with $N-p-1$ degrees of freedom where N is the sample size and p is the number of parameters included in the model. On examining the residuals I found that they had a much narrower variance than expected and thus the threshold of 0.01 was much more stringent that I was anticipating. This also explains why for ~12,000 individuals I only had a few individuals pass the threshold. On fitting a Poisson GLM, as I have done in other parental age analyses in this chapter, I found the variance of these studentized pearson residuals was inflated and so may also not be the correct approach. In my next steps I aim to improve this methodology (for example using quasi-Poisson or negative binomial regression) to ensure I am using the most appropriate model and capturing all possible hypermutated individuals. Although I would note that the rank order of hypermutated individuals is barely altered under these different models, only the p values change.

I found that germline hypermutation explained 8% of the variance in germline mutation rate in 100kGP. The fact that this is evaluated in a cohort that consists of offspring with genetic disorders may mean this is an overestimate of how much variance is explained by hypermutation in the general population. *De novo* mutations are a major cause of DD and cohorts of children with developmental disorders are enriched for DNMs overall and so would be more likely to contain hypermutated individuals [41]. In a healthy population this variance explained may be smaller. However we would still expect to see hypermutation in a healthy population. The absolute risk of a germline hypermutator having a child with a genetic disease is still low. The population average risk is estimated to be 1 in 300 births and so a 4 fold increase in DNMs in a child will amount to the risk of a genetic disease is just over 1% [41]. I found that parental age explained ~70% of the germline mutation rate variance which is substantially smaller than a previous estimate of 95% [111] based on a sample of families ~100x smaller than the one I analysed. This may be due to several factors such as differences in measurement error, non-random selection of parental age or by chance. Another possible contributing factor may be that in 100kGP the variance of the number of DNMs is larger than it would be in a healthy population. The remaining ~20 % of germline mutation variation remains unexplained in this analysis. Part of this may be attributable to additional hypermutated individuals that may be identified upon improving my model although this is unlikely to amount to a substantial additional fraction. Rare coding and non-coding variants in DNA repair genes or genes currently not known to be associated with germline mutation rate may also explain more variance. However even with thousands of whole genome sequenced trios we may not be powered to identify these across the genome. Another source of variation may be explained by polygenic effects on germline mutation rate. Previous work has demonstrated that there are differences in germline mutation rate between populations and that there are loci in the genome that may be associated with a higher germline mutation rate [75, 197]. A genome wide association study (GWAS) approach using the DNMs as a proxy for germline mutation rate in the parents requires parent-offspring trio sequencing just to measure the phenotype. This means sequencing 3x as many individuals as you expect to test which is costly especially considering that a very large sample size would be needed. Another possibility may be to conduct an association study on male germline mutation rate by using estimates of individual mutation rates from single cell sequencing of sperm. This would also allow interrogation of the within variation of individual mutation rate. This may be feasible as single cell technology and methodology improves and sequencing costs decrease however large sample sizes would be needed and a similar interrogation of the female germline mutation rate would not be feasible. Environmental effects are also likely to contribute to germline mutation rate variation so including deep phenotyping and details of possible exposures would be important to include in a large germline mutation rate study and may also help reveal gene by environment interactions.

The analyses in this chapter provide new insights into the role of genetic variation on the human germline mutation rate. I have demonstrated the existence of germline hypermutators as well as possible genetic causes. I have shown that hypermutation explains a significant proportion of germline mutation rate variation in addition to parental age but also that there is residual variance that still needs to be explored.

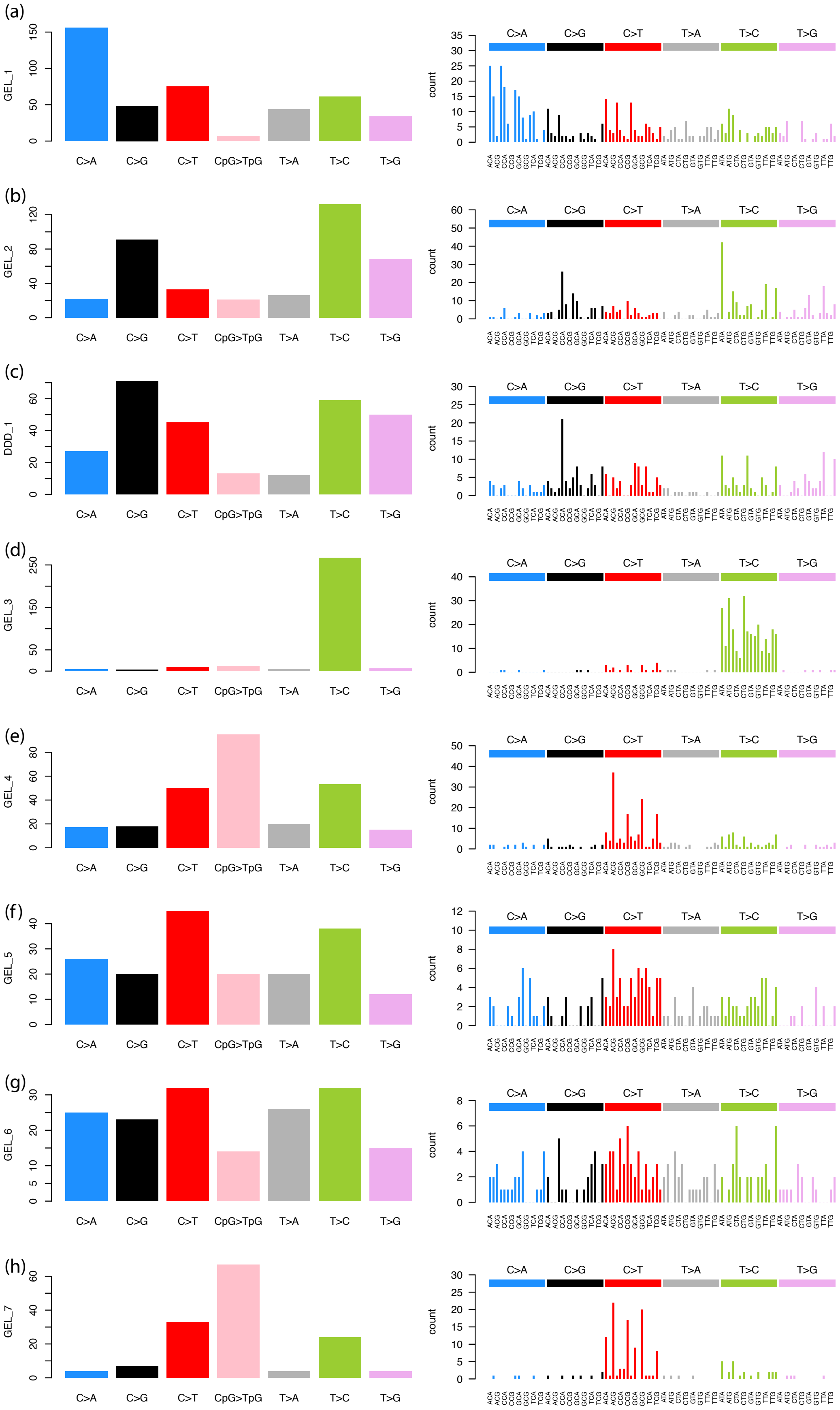


Fig. 3.14 Mutational spectra of DNMs from hypermutated individuals (A)

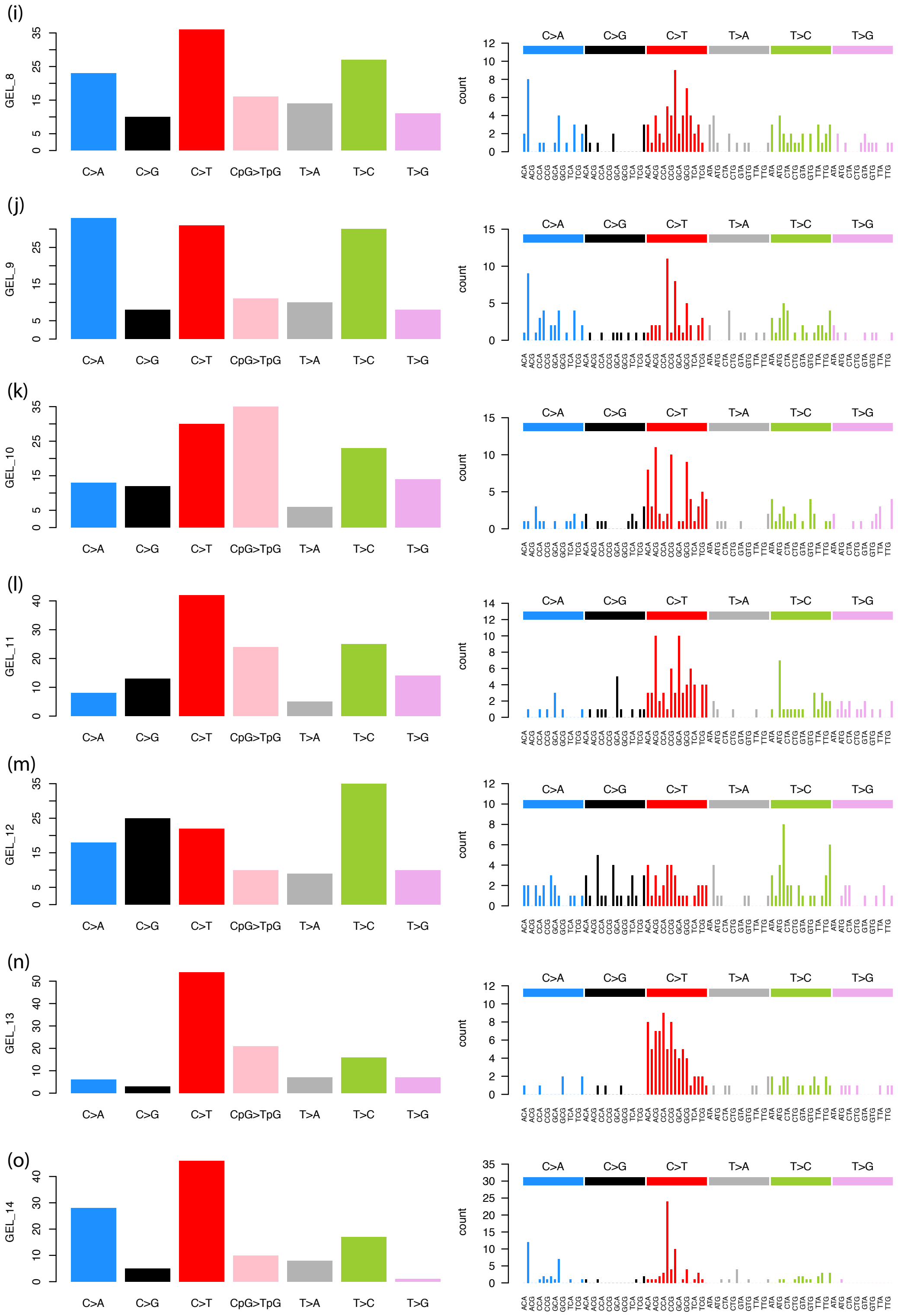


Fig. 3.15 Mutational spectra of DNMs from hypermutated individuals (B)

Chapter 4

Integrating healthcare and research genetic data empowers the discovery of 28 novel developmental disorders

4.1 Introduction

It has previously been estimated that 42-48% of patients with a severe developmental disorder (DD) have a pathogenic *de novo* mutation (DNM) in a protein coding gene [41, 140]. However, over half of these patients remain undiagnosed despite the identification of hundreds of dominant and X-linked DD-associated genes. This implies that there are more DD-associated genes left to find.

Whole genome and exome sequencing allows for direct identification of DNMs and statistical assessment of their contribution to DD. Associated genes are typically identified by observing a gene-specific enrichment of DNMs over what is expected by chance. Initially, when cohorts comprised of fewer than 200 trios, candidate DD-associated genes were proposed by identifying those with multiple non-synonymous DNMs across the cohort [157, 163, 58]. Improvements in the modelling of the mutation rate and increased sample sizes allowed for a more statistical approach. The most recent approaches taken in DD and autism cohorts have generally identified associated genes by separately evaluating the enrichment of protein truncating and missense DNMs compared to gene-specific mutation rates that account for gene length and sequence context[187, 215, 41, 31, 162]. This enrichment was calculated in some cases by assuming a Poisson model for the distribution of mutations and calculating this enrichment analytically[41, 215]. Other approaches were simulation based where a null distribution was calculated by simulating the observed number of mutations across all genes using a multinomial distribution[31, 162]. McRae et al, in a previous paper from the Hurles group evaluating the role of DNMs in ~4,000 trios from the DDD study, also combined the missense enrichment test with a missense clustering test within genes. Clustering of mutations within protein structures is frequently observed for DNMs acting via activating or dominant negative mechanisms. Statistical testing for clustered mutations was initially developed in the study of somatic mutation enrichment in cancer[119, 172]. Somatic mutation rate variability can lead to false positive associations if locus-specific mutation rates are higher than expected. Methods have been developed to protect against this which combine the local observed synonymous mutation rate with a model including variables that predict the somatic mutation rate across the genome[141]. Existing methods to detect gene-specific enrichments of damaging DNMs typically ignore much prior information about which variants and genes are more likely to be disease-associated. Missense variants and protein-truncating variants (PTVs) vary in their impact on protein function [107, 186, 102, 112]. A study in a cohort of neurodevelopmental disorders observed that *de novo* PTVs that do not appear as standing variation in healthy population cohorts and fall in genes that exhibit patterns of strong selective constraint on heterozygous PTVs in the general population contribute to the majority of the enrichment of *de novo* PTVs in the cohort[112]. Known dominant DD-associated genes are also strongly enriched in this set of PTV constrained genes[125]. To identify the remaining DD genes, we need to increase our power to detect gene-specific enrichments for damaging DNMs by both increasing sample sizes and improving our statistical methods. In previous studies of pathogenic Copy Number Variation (CNV), utilising healthcare-generated data has been key to achieve much larger sample sizes than would be possible in a research setting alone [36, 32].

4.1.1 Chapter overview

To identify novel DD-associated genes, I integrated healthcare and research exome sequences on 31,058 DD parent-offspring trios. These were pooled from the Deciphering Developmental Disorders study, GeneDx (a US-based genetic diagnostic company) and Radboud University Medical Center. I developed a simulation-based statistical test to identify gene-specific enrichments of DNMs. Applying this to the dataset I identified 285 significantly DD-associated genes, including 28 not previously robustly associated with DDs. I then explored how these genes differed to those previously known to the field and examined the remaining burden of DNMs in genes yet to be associated with DD. Despite detecting more DD-associated genes than in any previous study, much of the excess of DNMs of protein-coding genes remains unaccounted for. To address this I built a model to estimate how many genes are left to be discovered. The model suggests that over 1,000 novel DD-associated genes await discovery, many of which are likely to be less penetrant than the currently known genes.

4.1.2 Publication and contributions

The results described in this chapter have been submitted to the biorXiv preprint server and has recently been accepted to *Nature* [101]. This work was conducted as a collaboration between our group at the Wellcome Sanger Institute, GeneDx and Radboud University Medical Center (RUMC). I briefly summarise the various contributions to this project. Zhancheng Zhang called *de novo* mutations in GeneDx data, Kevin J. Arvai and Rebecca Torene performed the phenotypic comparison work, Stefan H. Lelieveld called and filtered *de novo* mutations in RUMC. Kaitlin Samocha (post-doc in the Hurles group) and I worked jointly on many aspects of this project. In this chapter I have mostly included analyses that I performed and otherwise have been explicit about who performed them. All of this work was done under the supervision of Christian Gillissen (RUMC), Kyle Retterer (GeneDx) and Matthew E. Hurles.

4.2 Methods

4.2.1 Sample collection and individual quality control

DDD

Patients with severe, undiagnosed developmental disorders were recruited from 24 regional genetics services within the United Kingdom National Health Service and the Republic of Ireland. Families gave informed consent to participate, and the study was approved by the UK Research Ethics Committee (10/H0305/83 granted by the Cambridge South Research Ethics Committee, and GEN/284/12 granted by the Republic of Ireland Research Ethics Committee). Additional details on sample collection, exome sequencing, alignment, variant calling (inherited and *de novo*) and variant annotation have been described previously [41]. These analyses involve 9,858 trios from 9,307 families, a subset of whom have been analyzed in previous publications [215, 41].

GeneDx

Patients were referred to GeneDx for clinical whole-exome sequencing for diagnosis of suspected Mendelian disorders as previously described[177]. Patient medical records were abstracted into HPO terms using Neji concept recognition[21] with manual review by laboratory genetic counselors or clinicians. Patients were selected for inclusion in this study based on having one or more HPO phenotypes overlapping the inclusion criteria for the DDD study [215]. The study was conducted in accordance with all guidelines set forth by the Western Institutional Review Board, Puyallup, WA (WIRB 20162523). Informed consent for genetic testing was obtained from all individuals undergoing testing, and WIRB waived authorization for use of de-identified aggregate data. Individuals or institutions who opted out of this type of data use were excluded.

The sequencing and variant calling was done by collaborators at GeneDx. The exomes were sequenced and aligned as previously described [177] with either SureSelect Human All Exon v4 (Agilent Technologies, Santa Clara, CA), Clinical Research Exome (Agilent Technologies, Santa Clara, CA), or xGen Exome Research Panel v1.0 (IDT, Coralville, IA) and sequenced with either 2x100 or 2x150bp reads on HiSeq 2000, 2500, 4000, or NovaSeq 6000 (Illumina, San Diego, CA). Alignment BAM files were then converted to CRAM format with Samtools version 1.3.1 and indexed. Individual GVCF files were called with GATK v3.7-0 HaplotypeCaller [143, 43] in GVCF mode by restricting output regions to plus/minus 50bp of the RefGene primary coding regions. Single-sample GVCF files were then combined into multi-sample GVCF files with each combined file contained 200 samples. These multi-sample GVCF files were then joint-genotyped using GATK GenotypeGVCFs. The cohort of 18,789 trios was joint-genotyped in two separate batches, one with 10,138 trios and the other 8651 trios. GATK VariantRecalibrator (VQSR) was applied for both SNPs and INDELs, with known SNPs from 1000 Genomes phase 1 high confidence set and “gold standard” INDELs from Mills et al [149].

Variants in VQSR VCF files were annotated with Ensembl Variant Effect Predictor (VEP)[144] using RefSeq transcripts. The transcript with the most severe consequence was selected, and all associated VEP annotations were based on the predicted effect of the variant on that particular transcript. Variants called in the proband and not in the parents were selected as potential *de novo* mutations. Filtering of these *de novo* mutations is described below.

Radboud University Medical Center

The Department of Human Genetics from the Radboud University Medical Center (RUMC) is a tertiary referral center for clinical genetics. Approximately 350 individuals with unexplained intellectual disability (ID) are referred annually to the clinic for diagnostic evaluation. Since September 2011 whole exome sequencing (WES) is part of the routine diagnostic work-up aimed at the identification of the genetic cause underlying disease[158]. For individuals with unexplained ID, a family-based WES approach is used which allows the identification of *de novo* mutations (DNMs) as well as variants segregating according to other types of inheritance, including recessive mutations and maternally inherited X-linked recessive mutations in males [39]. For this study, RUMC selected all individuals with ID who had family-based WES using the Agilent SureSelect v4 and v5 enrichment kit combined with sequencing on the Illumina HiSeq platform in the time period 2013-2018. This selection yielded a set of 2418 individual probands, including 1040 females and 1378 males across 2387 different families. The level of ID ranged between mild (IQ 50-70) and severe-profound (IQ<30).

Families gave informed consent for both the diagnostic procedure as well as for forthcoming research that could result in the identification of new genes underlying ID by meta-analysis, as presented here.

The sequencing and variant calling was done by collaborators at RUMC. The exomes of 2418 patient-parent trios were sequenced, using DNA isolated from blood, at the Beijing Genomics Institute (BGI) in Copenhagen. Exome capture was performed using Agilent SureSelect v4 and v5 and samples were sequenced on an Illumina HiSeq 4000 instrument with paired-end reads to a median target coverage of 112x. Sequence reads were aligned to the hg19 reference genome using BWA version v0.7.12 and duplicate marking by Picard v1.90. Variants were subsequently called by the GATK haplotypecaller (version v3.4-46).

The diagnostic WES process as outlined above only reports (*de novo*) variants that can be linked to the individuals' phenotype. In this study, we systematically collected all DNMs in regions in or close to (200 bp) a capture target. DNMs were called as described previously [39]. Briefly, variants called within parental samples were removed from the variants called in the child. For the remaining variants pileups were generated from the alignments of the child and both parents. Based on pileup results variants were then classified into the following categories: "maternal (for identified in the mother only)", "paternal (for identified in the father only)", "low coverage" (for insufficient read depth in either parent), "shared" (for identified in both parents)", and "possibly *de novo*" (for absent in the parents). Variants classified as possibly *de novo* were included in this study.

Various quality filters were applied to ensure that only the most reliable calls were included in the study, these are described in the quality control section below.

4.2.2 Definition of diagnostic lists

For various analyses in this work, a list of genes already known to be associated with developmental disorders was needed. In order to define this list, diagnostic gene lists were collected from each center to create sets of "consensus" and "discordant" genes.

For the DDD cohort, the Developmental Disorders Genotype-Phenotype Database (DDG2P) list was used. This is a curated list of genes specifically associated with developmental disorders. For every gene on the list, DDG2P provides the level of certainty, consequence of the mutation, and allelic status of variants associated with developmental disorders. We downloaded the DDG2P list on 22 September 2019 from https://decipher.sanger.ac.uk/info/ddg2p. In order to define diagnostic genes that act in a dominant fashion, the genes were subsetted to include only genes that were considered "probable", "confirmed", or "both RD and IF" (i.e. high levels of certainty of being a true DD-associated gene) and had an allelic status of "monoallelic", "x-linked dominant", "hemizygous", or "imprinted".

GeneDx maintains a continually curated list of genes, used to define reporting categories for clinical exome and genome testing, which have been definitively or putatively implicated in human Mendelian disease, with modes of inheritance noted for each gene. Starting with the January 2020 curation list, those genes with dominant modes of inheritance and definitive implications in disease were manually reviewed to remove any genes with no association to developmental disorders either because of no phenotypic overlap with the inclusion criteria for this study or because the relevant phenotypes were adult onset.

For the list from RUMC, gene panels for intellectual disability, epilepsy, and craniofacial anomalies/Multiple congenital anomalies were designed by multidisciplinary expert teams consisting of a clinical laboratory geneticist, a molecular geneticist, and a clinical specialist. Each set contained all genes known to be associated with the disease. The gene panel version from December 2019 (DGD-2.17) was used. From each of the three gene lists, the genes were subsetted to those with a reported inheritance of “AD”, “AD,AR”, “AD,IMP”, “AD/AR”, “xl”, “XL”, “XLD”, “XLR,XLD”, or “XLR/XLD”.

After mapping to HGNC IDs and symbols, any gene that was considered diagnostic by all three centers was designated as a “consensus” gene (n=380). For genes on one or two of the diagnostic lists, we considered them “discordant” genes (n=607).

4.2.3 Joint quality control of datasets

De novo mutation filtering

I applied the following filters specifically to each center. I chose these to minimise the number of false positive DNMs. I evaluated this by looking at various plots and metrics. These included plots of the variant allele fraction across each cohort to ensure this was centred symmetrically around 0.5, the distribution of the number of exome DNMs per person, the mutational spectra of the DNMs, the distribution of the size of indels and the ratios of insertion/deletion, frameshift/nonsense and frameshift/inframe variants. I compared these across each cohort to ensure they were comparable. The VAF distributions pre and post filtering VAF distirbutions are shown in Figure 4.1 . The DNMs from the DDD dataset in the 'pre-filtering' had already undergone a basic set of the filters described below such as read depth, strand bias and number of parental alt alleles. The DNMs from RUMC and GeneDx had undergone prior filtering which are specified below. The filters were applied in this specific order.

DDD:

- Autosomes
  - I applied the following filters as base filters before attempting to harmonize the three datasets (these are applied to the 'pre-filtering' set as shown in Figure 4.1).
    - The minor allele frequency (MAF) < 0.01 across all DDD samples, Exome Aggregation Consortium (ExAC)[125], and 1000 Genomes[34] populations
    - Read depth (RD) of child > 7, mother RD > 5, father RD > 5
    - Fisher exact test on strand bias p-value > 10^{-3}
    - Remove DNM if any two of the following conditions are met:
    - Both parents had ≥ 1 supporting the alternative allele
    - There is an excess of parental alternative allele within the cohort at the DNMs position. This is defined as p-value $< 10^{-3}$ under a one-sided binomial test given an expected site error rate of 0.002
    - There is an excess of alternative alleles within the cohort for DNMs in a gene. This is defined as p-value < 10^{-3} under a one-sided binomial test given an expected site error rate of 0.002
  - Filter only applied to indels; remove indel if all three conditions are met:
    - Variant allele frequency (VAF) in child < 0.2
    - MAF > 0 for any of the following cohorts: across all DDD samples, ExAC, 1000 Genomes populations
    - Size of indel < 5 bp
  - Posterior probability of being a *de novo* mutation (output from DeNovo Gear) > 0.00781 for autosomal DNMs. These thresholds have been determined through earlier work such that the observed number of synonymous DNMs match the expected number [41, 140].
  - Filter out mutations in sites with more than one mutation with VAF < 0.3
- X chromosome: DeNovoGear was run as previously described, but with a different set of hard filters to account for the lower coverage in males and to maximise sensitivity

and specificity. These were developed by Hilary Martin. All candidate DNMs in males and a large subset of those in females were inspected manually in IGV[182], and this was used to settled on the following set of filters:

- Removed DNMs in the pseudoautosomal regions.
- The variant had to be called heterozygous or, for males, hemizygous in the child in the original GATK calls, and called homozygous reference in the parents.
- Removed variants in segmental duplications.
- For male probands, the depth requirements were as follows: in the child, alternate allele depth > 2 and RD > 2; in the mother, RD > 5
- For female probands,the depth requirements were as follows: in the child, alternate allele depth > 2, RD > 7; in the mother, RD > 5; in the father, RD > 1
- For single nucleotide variants, a $p>10^{-3}$ was required on a Fisher's exact test for strand bias, pooling across trios (ignoring fathers of male probands) where a *de novo* was called at the same site by DeNovoGear.
- For female probands, indels < 5 bp were removed if they had VAF < 0.3 or MAF > 0, since these were vastly over-represented and seemed to be a common error mode.
- Removed DNMs if any two of the following conditions were met (these conditions were applied separately for males and females):
    - Lowest alternative read count for the parents (or, for males, the mothers) is higher than the maximum allowable given the depth, an error rate of 0.002, and a probability threshold of 0.98 (using the mindepth function in DeNovoFilter)
    - An excess of parental (or, for males, maternal) alternative alleles with a putative *de novo* at that site, defined as p-value $< 10^{-3}$ under a one-sided binomial test given an expected error rate of 0.002
    - An excess of parental (or, for males, maternal) alternative alleles with a putative *de novo* in the same gene, defined as p-value $> 10^{-3}$ under a one-sided binomial test given an expected error rate of 0.002
- Implemented a cutoff for the ppDNM from DeNovoGear to > 0.00085 based on matching the observed number of synonymous DNMs in females to the expected number

GeneDx:

- The following filters were initially applied by collaborators at GeneDx:
    - RD > 10 for child, mother, and father
    - VAF > 0.15 for child for SNVs and VAF > 0.25 for indels
    - More than 3 reads supporting the alternative allele
    - Genotype Quality (GQ) score > 40
    - Phred-scaled p-value using Fisher's exact test to detect strand bias < 30
    - Log odds of being a true variant versus being false from VQSR > -10 outputted from GATK
    - Any variant with general population frequency above 0.01 was also excluded based on 1000 Genomes and Exome Aggregation Consortium (ExAC) variant population frequency data
    - Filtered out *de novo* variants called > 4 times in the parental samples in the cohort
- Filter out DNMs with VAF < 0.3 and VQSLOD < 7 where VQSLOD is the log odds ratio of being a true variant outputted from GATK
- Filter out *de novo* indels > 100 bp
- Filter out DNMs, not on chromosome X, with a VAF of 1
- Filter out 5 individuals with more than 10 coding DNMs. These appeared to be due to relatively poor sample quality.

RUMC:

- The following filters were initially applied by collaborators at RUMC:
    - Minimal number of variant reads: 10
    - Minimal number of total reads: 20
    - Minimal percentage of variant reads: 20%
    - Frequency in dbSNP < 0.1%
    - Coverage in parents of at least 10 reads
    - 15 complex variants were discarded after manual inspection in IGV
- GATK Quality score > 450

The following filters were applied across all three datasets:

- Removed DNMs outside coding regions
- Removed mutations that fell within known segmental duplication regions as defined by UCSC (http://humanparalogy.gs.washington.edu/build37/data/GRCh37GenomicSuperDup.tab)
- Keep only the most severe DNM in each gene per individual according to the consequence severity from VEP[144]

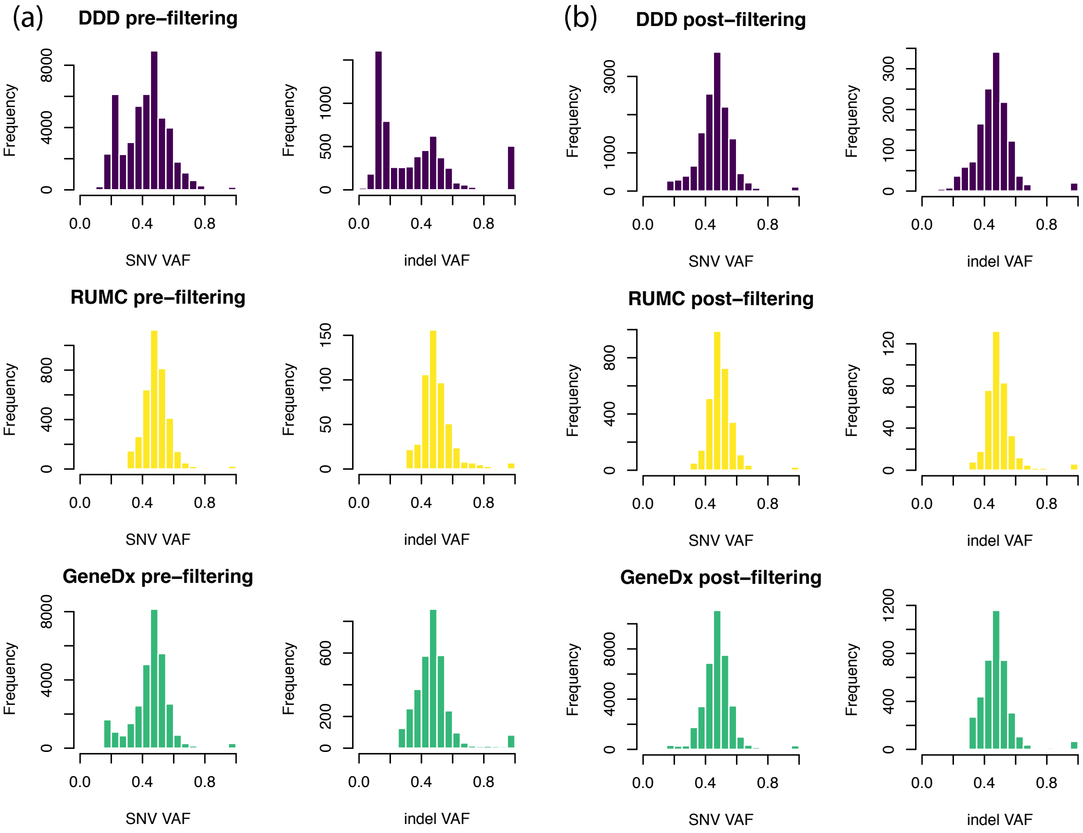


Fig. 4.1 Variant allele fraction (VAF) for all *de novo* coding mutations, split by center pre and post filtering. The pre-filtering sets did include base filters as specified in the Methods. Note that mutations with a VAF of 1 are hemizygous *de novo* mutations in males. DDD = Deciphering Developmental Disorders. RUMC = Radboud University Medical Center.

Duplicate Samples

50 common exonic SNPs were selected (only 47 of which could be evaluated across all three centers) and collected genotypes at these SNPs for every sample with a *de novo* mutation

found in another individual in the joint set (n=781 DDD, 1307 GeneDx, and 164 RUMC). Kaitlin Samocha then used the gtcheck function from bcftools (https://samtools.github.io/bcftools/bcftools.htm to find discordance between each pair of samples. Pairs with low discordance were manually confirmed, leading to a total of 8 duplicate samples identified. One individual from each duplicate pair was removed from the analyses, leaving 31,058 samples for downstream analyses.

Removing variants from siblings

Siblings will sometimes share DNMs that arose as the same mutational event in one of their parental germlines. To avoid double counting these shared DNMs, Kaitlin identified siblings in each cohort and one of each pair of shared variants was randomly removed. In total, 11 DNMs found in siblings were removed.

After filtering I had a total set of 45,221 coding DNMs (these are available online as supplementary table 1 with the bioRxiv preprint [101]). The filtered VAFs across the three datasets are displayed in Figure 4.1b.

4.2.4 DeNovoWEST framework

DeNovoWEST (*De novo* Weighted Enrichment Simulation Test) is the testing framework I developed and applied to assess gene-wise *de novo* mutation enrichment. Each observed DNM in our dataset was assigned a mutation severity score. This severity score is a proxy for how deleterious we expect the mutation to be. Details of how these were calculated are given below. For each gene I then calculated a gene severity score which is the sum of severity scores for all mutations that fall into that gene. There are two modes of enrichment testing within DeNovoWEST: the overall enrichment test which includes all variant consequences and the ’altered-function’ specific test which assesses enrichment and clustering of missense variants only. An overview of the method is given in Figure 4.2.

Enrichment Test

I used a simulation-based approach to evaluate whether these observed gene severity scores are higher than what we would expect under the null hypothesis of no *de novo* mutation enrichment. To calculate the probability of observing a gene severity score that is as or more extreme than the one that we observe for this gene (the p-value) we considered the case of observing k number of DNMs in the gene where k ranged from 0 to 250. This upper limit was chosen as it is far above the number of DNMs seen in any individual gene in the dataset and the probability of observing more than that number of DNMs for our cohort in a single gene

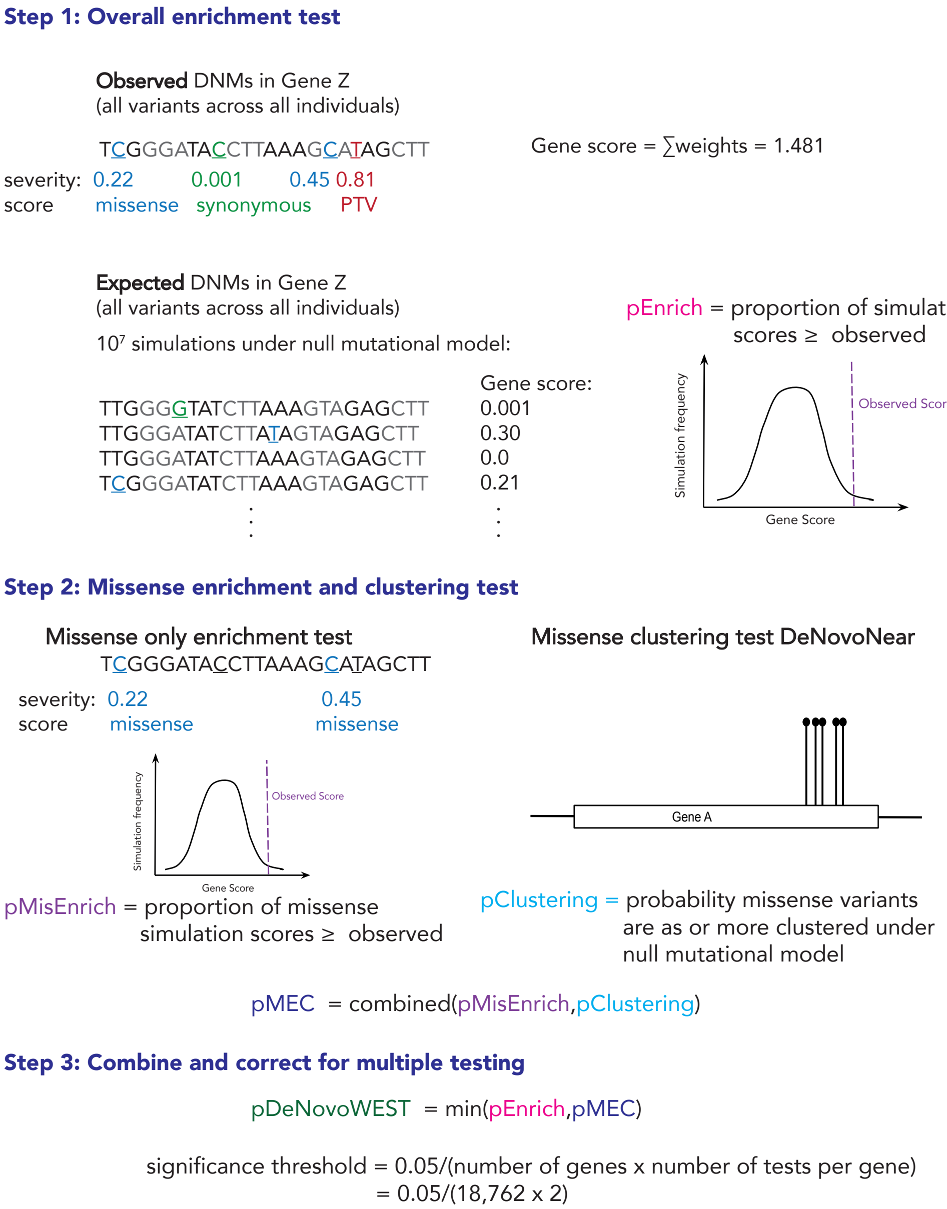

Step 1: Overall enrichment test
Observed DNMs in Gene Z
(all variants across all individuals)
TCGGGATACCTTAAAGCATAGCTT
Gene score = ∑weights = 1.481
severity: 0.22 0.001 0.45 0.81
score missense synonymous PTV
Expected DNMs in Gene Z
(all variants across all individuals)
10^7 simulations under null mutational model:
Gene score:
TTGGGGTATCTTAAAGTAGAGCTT 0.001
TTGGGATATCTTATAGTAGAGCTT 0.30
TTGGGATATCTTAAAGTAGAGCTT 0.0
TCGGGATATCTTAAAGTAGAGCTT 0.21
pEnrich = proportion of simulation scores ≥ observed
Simulation frequency
Observed Score
Gene Score
Step 2: Missense enrichment and clustering test
Missense only enrichment test
TCGGGATACCTTAAAGCATAGCTT
severity: 0.22 0.45
score missense missense
Simulation frequency
Observed Score
Gene Score
pMisEnrich = proportion of missense simulation scores ≥ observed
Missense clustering test DeNovoNear
Gene A
pClustering = probability missense variants are as or more clustered under null mutational model
pMEC = combined(pMisEnrich,pClustering)
Step 3: Combine and correct for multiple testing
pDeNovoWEST = min(pEnrich,pMEC)
significance threshold = 0.05/(number of genes x number of tests per gene)
= 0.05/(18,762 x 2)

Fig. 4.2 Overview of DeNovoWEST method

is negligible with respect to our significance threshold. The enrichment p-value, pEnrich, was then calculated as the sum across all k of the product of the probability of observing k mutations and the probability of obtaining a gene severity score greater than the one observed in our data. These probabilities are summarized by the following equation where S denotes the gene score, s is the observed gene score and K is the number of DNMs in the gene:

$$\begin{aligned}P(S \geq s) \approx &\sum_{k=0}^{250} P(S \geq s|k)P(K=k)\\ =&P(K=0)P(S \geq s|K=0) + P(S \geq s|K=1)P(K=1) + \sum_{k=2}^{250} P(S \geq s|k)P(K=k)\end{aligned}$$

These probabilities were calculated as follows:

- $P(S \geq s|K=0)$ = 0 (unless $s = 0$ in which case we would not be testing this gene as it would have no observed DNMs).
- I also analytically calculated $P(S \geq s|K=1)P(K=1)$, the probability that the severity score of 1 mutation was greater than what we observed. This was calculated using the mutability of each position and the annotated score for that mutation.
- I then used a simulation-based approach to calculate $\sum_{k=2}^{250} P(S \geq s|k)P(K=k)$, the probability of observing an as, or more, extreme gene score if we see 2 to 250 DNMs in this gene. I calculated $P(K=k)$ analytically assuming the DNMs followed a poisson distribution. This was calculated based on our sample size (N) and the mutation rate of the gene assuming mutations follow a Poisson distribution with $\lambda = 2\mu_{gene}N$. This was adjusted to $\lambda = \mu_{gene}(N_{males} + 2N_{females})$ for genes that fall in the X chromosome. I then multiplied this with a simulation based estimate for $P(S \geq s|k)$, the probability of observing a gene score greater or equal to the one we observe. To calculate the latter probability I simulated the distribution of gene scores as follows:
  1. Simulate k DNMs across the gene. The probability of mutation was weighted by the trinucleotide sequence context at every base position in that gene. These probabilities were taken from Samocha et al [187].
  2. Assign the simulated *de novo* mutations a mutation severity score
  3. Sum the simulated mutation severity scores to get the simulated gene severity score
  4. I performed $10^9 \times P(K=k)$ simulations for every k. This number was chosen as I wanted to run the smallest number of simulations possible to obtain a robust p-value. By distributing these simulations across the mutations this was the equivalent of 10^9 simulations per gene and meant that the p-value was robust to stochasticity far below the p-value threshold.

Determination of Weights

I calculated the weights used in the DeNovoWEST test from observed enrichments across mutation consequence classes, s_{het} values, missense constraint information and, for some, CADD score bins (version 1.0)[107]. s_{het} refers to the estimated selection coefficient of heterozygous PTVs [24]. Genes were stratified into two groups of 'high' s_{het} and 'low' s_{het}. High s_{het} genes were defined as those with a $s_{het} \geq 0.15$ and low s_{het} genes as those with an estimate <0.15. This threshold was suggested by Cassa et al. Enrichments were calculated by dividing the number of observed *de novo* mutations by the number of expected *de novo* mutations across all sites in the exome that fell into a specific strata. I calculated the number of expected mutations given our sample size and the triplet context mutation rates at the sites [187]. Details for the weight calibration for each consequence class are given below:

- Missense mutations were stratified based on whether they fell in a low or high s_{het} gene [24], whether or not they fell into a region of missense constraint [186], and finally into CADD score bins of size 6[107]. I fit four LOESS lines on the enrichments for mutations that were in high s_{het} genes + missense constrained regions, high s_{het} genes + not in a missense constrained region, low s_{het} genes + missense constrained regions, low s_{het} genes + not in a missense constrained region.

- Nonsense mutations were stratified based on whether they fell in a low or high s_{het} gene, and then into CADD score bins of size 15 for the high s_{het} genes and CADD score bins of 7.5 for low s_{het} genes. Two LOESS lines were fit on the enrichments for mutations that were in high s_{het} genes vs those in low s_{het} genes.

- Synonymous mutations were stratified based on whether they fell in a low or high s_{het} gene.

- Canonical splice site mutations were stratified based on whether they fell in a low or high s_{het} gene.

- Inframe indels were assigned weights based on the overall enrichment of missense mutations as an appropriate approximation for their deleteriousness. These were stratified by whether they fell in a low or high s_{het} gene but not stratified by CADD score bins.

- Frameshift indels were assigned the same weights as nonsense mutations with a CADD score ≥ 45 and whether they fell in a low or high s_{het} gene.

These enrichments are depicted in Figure 4.3. The enrichment values for each stratum were normalised by the level of synonymous enrichment and converted into a positive predictive value (PPV) using the following formula:

$$PPV = \frac{OR - 1}{OR} \tag{4.1}$$

The synonymous variants were artificially given a PPV of 0.001 as we would expect 1 in 1000 synonymous DNMs in our cohort to be pathogenic according to how many are estimated to be cryptic splice sites[91]. The PPV weights are depicted in Figure 4.3.

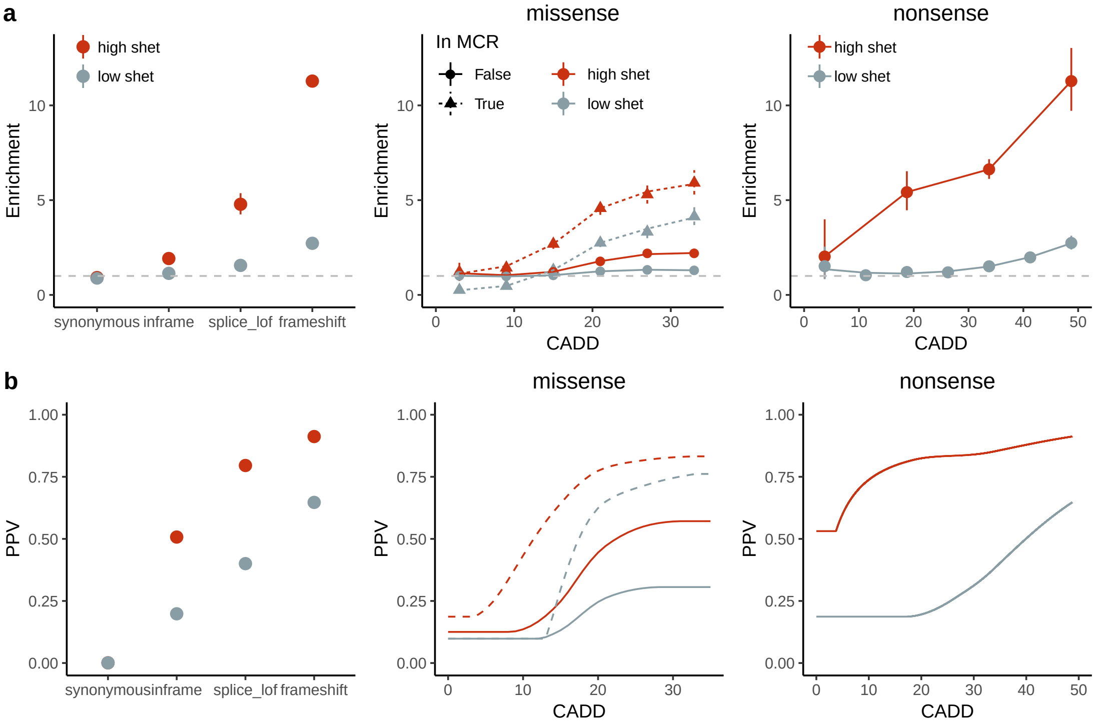


Fig. 4.3 Enrichment of consequence classes and corresponding PPV weights used for DeNovoWEST test. (a) Depicts the observed enrichment in each consequence class with 95% confidence intervals. The lines fit in the missense and nonsense class are LOESS fits. (b) Depicts the PPV derived from these observed enrichments. In all these plots points and lines are colored red if the variants occurred in a gene with a high s_{het} value (≥ 0.15)[24] or gray if the gene had a s_{het} value < 0.15 ("low s_{het}" genes). For missense variants, dashed lines indicate that the variants fell within missense constrained regions (MCR) while solid lines fell outside of MCRs.

Missense enrichment and clustering test

This test is geared to detect genes that may be acting via an altered-function mechanism, such as a gain-of-function. The test consists of two parts. The first is implementing the enrichment test (as described above) but only considering missense variants to obtain a pMisEnrich p-value. The second part consists of a missense clustering test. I assessed clustering of missense *de novo* mutations within genes to identify genes where DNMs may be acting through dominant negative or activating mechanisms. For this I used DeNovoNear, which has been described previously [41] and is available on Github (https://github.com/jeremymcrae/denovonear). I refer to this clustering p-value as pClustering. I combined pMisEnrich and pClustering using Fisher's method to obtain pMEC. To ensure Fisher's method was appropriate I confirmed that these two p-values were independent by simulating DNMs under the null for ~60,000 genes and found a nonsignificant correlation between pMissenseEnrich and pCluster (ρ = -0.01, p-value 0.08).

Combining Tests

I combined the results from the overall enrichment test and the missense enrichment/clustering test by taking the minimum of the two p-values as follows:

$$pDenovoWEST = min(pEnrich, pMEC) \tag{4.2}$$

To correct for multiple testing, I used a Bonferonni corrected significance threshold of 0.05/(18762 * 2) for pDeNovoWest. This accounts for testing all 18,762 genes that I was able to conduct tests for and for 2 tests per gene: the overall enrichment test and the specific gain-of-function test. The final set of DeNovoWEST p-values are listed online as supplementary table 2 of the bioRxiv preprint [101].

4.2.5 Functional similarity between new and known genes

To compare the functional similarity between the consensus and novel genes I looked across various properties that have been known to be important in classifying haploinsufficiency[83]. The details of each variable we have used are detailed below:

- Somatic driver gene: a binary variable of whether the gene is a known somatic driver gene[141].
- Median reads per kilobase million (RPKM) fetal brain: the median RPKM in the fetal brain taken from BrainSpan[148].

- Relevant GO term: a binary variable of whether the gene was annotated with one of twenty GO terms that were enriched in consensus DD genes. To select these terms I annotated all genes with GO terms and looked at the enrichment of each GO term between consensus DD genes and non-DD genes (genes that are not significant in our analysis and are not on either the discordant or consensus genes lists). I defined relevant GO terms as the top 20 most enriched terms that appear in at least 20 of the 380 consensus genes. This was to ensure that I was picking terms that were generalisable to the entire set were not specific to only a few genes. The terms selected are detailed in Table 4.1. At least one of these 20 terms were present in 237 (71%) of consensus genes and 2,874 (16%) of non-DD genes.

- Network distance to consensus genes: As in Huang et al[83], Kaitlin created a protein-protein interaction network by integrating information from the Human Protein Reference Database[104], STRING[211], and Reactome[38]. Kaitlin then calculated the shortest path distance (a measure of proximity) between each gene and consensus genes.

- Network degree and betweenness: Kaitlin used MCL[220] (version 14-137; https://micans.org/mcl/) to determine network degree and betweenness (both measures of centrality).

- Promoter GERP and Coding GERP: These were calculated as described in Huang et al. [83]

- Macaque dN/dS: downloaded from Ensembl.

I compared the mean of these variables across consensus and novel genes compared to non-DD genes (Figure 4.7b).

4.2.6 DNM enrichment in non-significant genes

I calculated the remaining DNM burden in the genes that were not significantly associated with developmental disorders (DD) in our analysis and were not consensus DD genes. There were 2,172 genes that were not associated with DD and had a pLI $\geq$ 0.9 (high pLI) and 10,472 genes with a pLI < 0.9 (low pLI)[102]. The burden was calculated by calculating the observed/expected number of DNMs across the four groups of genes categorised by both missense/PTV mutations and low/high pLI. I then repeated the analysis removing nominally significant genes (unadjusted p-value > 0.05).

GO ID	GO name	Number of non-DD genes with GO term	Number of consensus DD genes with GO term	Enrichment of GO term in consensus vs non-DD genes
GO:0001501	skeletal system development	119	31	12.577748
GO:0007507	heart development	161	33	9.896377
GO:0007605	sensory perception of sound	107	21	9.47597
GO:0008543	fibroblast growth factor receptor signaling pathway	138	24	8.396926
GO:0001701	in utero embryonic development	217	37	8.23247
GO:0044212	transcription regulatory region DNA binding	168	28	8.047054
GO:0003682	chromatin binding	326	53	7.84958
GO:0010628	positive regulation of gene expression	155	24	7.475972
GO:0007411	axon guidance	295	44	7.201431
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	535	76	6.858797
GO:0043234	protein complex	323	42	6.278197
GO:0045893	positive regulation of transcription, DNA-dependent	512	66	6.223893
GO:0007268	synaptic transmission	359	45	6.052102
GO:0007420	brain development	200	25	6.03529
GO:0005667	transcription factor complex	220	27	5.925558
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	711	85	5.772148
GO:0007399	nervous system development	285	33	5.590585
GO:0003713	transcription coactivator activity	223	25	5.412816
GO:0008134	transcription factor binding	281	30	5.154696

Table 4.1 Table showing the GO terms selected as being relevant to consensus DD genes

4.2.7 Modelling remaining PTV DNM burden

I modelled the remaining DNM burden separately for PTV and missense mutations.

Model for PTV DNM burden

I simulated the number of *de novo* PTVs across a range of numbers of remaining haploinsufficient genes and the PTV enrichment observed in these genes. The PTV model made the following assumptions:

- PTV enrichment was the same across all remaining undiscovered HI DD-associated genes.
- All undiscovered HI DD-associated genes have the same level of penetrance
- The likelihood of being an undiscovered HI DD-associated gene was $1 - power$, where power was calculated as described below
- The probability of a currently unassociated gene being selected as a HI DD gene was higher if the gene was intolerant of loss-of-function variation. Specifically the likelihood was multiplied by the observed relative likelihood of being a DD-associated gene (significant in our analysis or a consensus gene) for genes with pLI ≥ 0.9. This was defined as:

$$\frac{P(DD\ associated\ gene|pLI \geq 0.9\ \&\ power > 0.8)}{P(DD\ associated\ gene|pLI < 0.9\ \&\ power > 0.8)} \tag{4.3}$$

- The PTV enrichment for known DD-associated genes (significant in the analysis or consensus) was taken as the observed enrichment in our cohort
- I ignore missense variants that may act as loss-of-function

Power as referred to in the above assumptions was calculated as follows. Calculating power tailored specifically according to our *de novo* enrichment test was challenging as I would have had to make assumptions about the distribution of mutation consequences according to undiscovered DD-associated genes. Even with these assumptions, the calculation would be computationally intensive given the simulation based framework. I decided to calculate power in the context of the enrichment of PTV mutations. Therefore this power measure is specifically the power to detect haploinsufficient genes. Power was defined as the power to detect the median PTV enrichment in known monoallelic DD-associated genes. The set of known monoallelic DD-associated genes was defined as 163 genes in the consensus gene list that had at least one observed *de novo* PTV in our dataset. I calculated the PTV enrichment by dividing the observed number of *de novo* PTVs in each gene by the expected number of *de novo* PTVs as defined by the gene-specific mutation rate. The median of the PTV enrichment distribution was 34.0. Power was then calculated as the probability of observing a significant p-value under the Poisson test assuming the median PTV enrichment.

To assess the likelihood of the model I calculated the following across the distribution of all 200 simulations per scenario:

$$\begin{aligned}Likelihood =&P(6,861\ PTV\ DNMs)P(147\ significantly\ PTV\ enriched\ genes)\\&\times P(2,929\ genes\ with\ PTV\ enrichment > 2)\end{aligned}$$

This captures three essential parts of the distribution of observed *de novo* PTVs: the total number of *de novo* PTVs, the number of genes that are currently significantly enriched for PTVs in our cohort and the number of genes that are not significant but have an elevated PTV enrichment (> 2). I explored other properties to characterise the distribution such as the number of genes with an enrichment p-value below a larger nominal threshold (eg 0.05) however the PTV enrichment appeared to be perform better. I evaluated this by examining distributions of PTV counts per gene and seeing if the most likely scenario using different metrics was similar to the observed distribution. Using this approach I also explored altering the threshold of PTV enrichment from a range of 1.5 to 5.

Model for missense DNM burden

The set up for the model for missense mutations was very similar to the PTV model with a few key differences. I simulated the number of *de novo* missense variants across a range of numbers of genes with a pathogenic DD-associated variant, mean missense enrichments in these genes and distributions of missense enrichment. I included a third dimension to the likelihood space which allowed me to model the distribution of missense enrichment across the genes with a pathogenic DD-associated missense variant.

I modelled the distribution of missense enrichment using the gamma distribution. I used 6 different shape parameters to represent different scenarios (0.1,0.5,1,510,20) (Figure 4.13 a). Missense mutations acting via gain-of-function mechanisms are likely to act on a small mutational target within a gene and have a small gene-wise missense enrichment. A shape parameter of 0.1 represents a model where most genes with pathogenic missense DNMs are acting via gain-of-function mechanisms and have generally smaller missense enrichment values. Missense mutations acting via loss-of-function mechanisms in haploinsufficient genes will have a larger mutational target and so the enrichment values will tend to be larger. A shape parameter of 20 represents the model where most genes with pathogenic missense variants are acting via loss-of-function and have larger enrichments. The other major difference from the PTV model was that the likelihood of being selected as a DD-associated gene in the simulations was not a function of pLI. It was only proportional to $1 - power$.

To assess the likelihood of the model I calculated the following across the distribution of all 200 simulations per scenario:

$$\begin{aligned} Likelihood = &P(27,139 \; missense \; DNMs)P(130 \; significantly \; missense \; enriched \; genes) \\ &\times P(3,764 \; genes \; with \; missense \; enrichment > 2) \end{aligned}$$

4.2.8 Expression in fetal brain

I defined expression in the fetal brain for a gene as having a median RPKM > 0 in the BrainSpan dataset[19]. I then compared the proportion of genes expressed in the fetal brain between the genes that were significant and not significant in our analysis . I then subsetted genes into those with a high pLI (pLI $\geq$ 0.9) and repeated this analysis.

4.3 Results

4.3.1 Improved statistical enrichment test identifies 300 significant DD-associated genes

Following clear consent practices and only using aggregate, de-identified data, DNMs in patients with severe developmental disorders were pooled from three centres: GeneDx (a US-based diagnostic testing company), the Deciphering Developmental Disorders study, and Radboud University Medical Center. I performed stringent quality control on variants and samples. I adjusted filters on read depth, VAF and *de novo* calling quality scores individually for each cohort. These were chosen to ensure the distribution of DNMs per person, mutational spectra and VAF distribution were consistent across the three datasets. After filtering I obtained 45,221 coding and splicing DNMs in 31,058 individuals (Figure 4.4), which includes data on over 24,000 trios not previously published. These DNMs included 40,992 single nucleotide variants (SNVs) and 4,229 indels.

All three cohorts are comprised of individuals with severe developmental disorders. The cohorts had comparable rates of male to female probands (55-57% male cohorts; Figure 4.4a) as well as similar DNM rates (average 1.81-1.96 per individual exome). The DDD study has a significantly higher rate of synonymous DNMs (0.31 *de novo* synonymous DNMs per exome) compared to individuals from GeneDx (GDX; 0.28 per exome; Poisson rate test p = 5.2×10^{-7}) or Radboud University Medical Center (RUMC; 0.28 per exome; Poisson rate test p = 0.0132), which is likely due to differences in *de novo* identification pipelines (Figure 4.4b). Specifically, as described in McRae et al[41] and mentioned in the methods, the DDD study selected a ppDNM (posterior probability of a *de novo* mutation) threshold such that the observed number of synonymous DNMs matched the expected number under a null germline mutation model.

All three cohorts have far more carriers of nonsynonymous DNMs in the consensus genes than expected based on a null mutational model (Figure 4.1c). The specific rate of such carriers differs between cohorts, with GeneDx showing the lowest fraction of such cases, which can be explained by the varying ascertainment between centers.

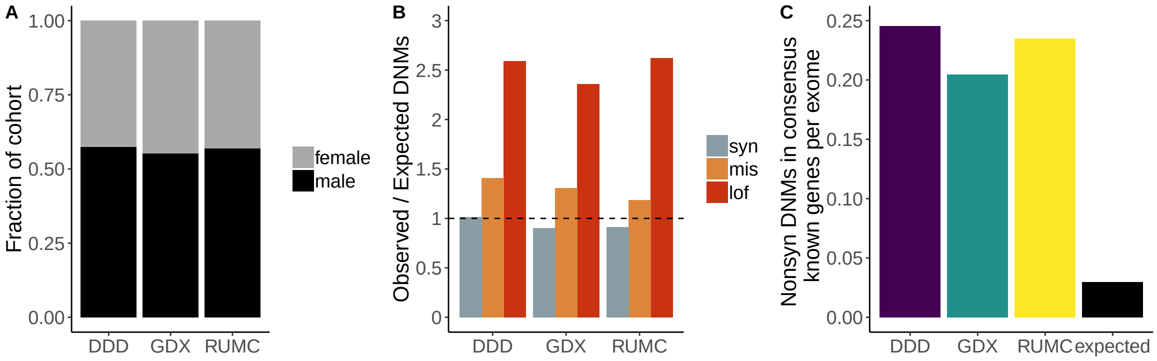


Fig. 4.4 Comparing cohorts from the three centers. A) Fraction of each cohort that is female (gray) vs male (black). B) Enrichment of observed *de novo* mutations compared to the expected number from a sequence-context based mutational model [187]. C) Rate of nonsynonymous *de novo* mutations (excluding inframe indels) in consensus genes in each cohort as well as the expected rate based on the aforementioned mutational model. DDD = Deciphering Developmental Disorders. GDX = GeneDx. RUMC = Radboud University Medical Center. syn = synonymous. mis = missense. lof = loss-of-function (including nonsense/stop gained, essential splice site, and frameshift variants).

To detect gene-specific enrichments of damaging DNMs, I developed a method named DeNovoWEST (*De Novo* Weighted Enrichment Simulation Test, described in detail in Methods; https://github.com/queenjobo/DeNovoWEST). DeNovoWEST assigns each observed DNM in our dataset a mutation severity score. This severity score is based on the empirically estimated positive predictive value of being pathogenic (Figures 4.2,4.3). For each gene the observed severity scores are then summed to obtain a gene score. Enrichment tests are then performed by comparing this observed gene score to a simulated null distribution. I performed two tests per gene: the first is an enrichment test on all nonsynonymous DNMs and the second is a test designed to detect genes likely acting via an altered-function mechanism. This second test combines an enrichment test on missense DNMs with a test of linear clustering of missense DNMs within the gene. I then applied a Bonferroni multiple testing correction accounting for 18,762 x 2 tests, which takes into account the number of genes and two tests per gene.

I first applied DeNovoWEST to all individuals in our cohort and identified 281 significant genes, 18 more than when using the method described in McRae et al (Figure 4.5a). I also ran DeNovoWEST on the DNMs from the ~4k DDD trios from the 2017 publication[41] and found an increase of 9 additional significant genes which again demonstrates a ~10% increase in power. (Figure 4.6a). The previous method consisted of a PTV enrichment test, a missense enrichment test and a missense clustering test. This previous method tested enrichment on counts of DNMs and variant consequence classes were treated separately.

As a negative control analysis, I applied DeNovoWEST to only synonymous DNMs . While synonymous mutations can be pathogenic, it is expected that, as a class, they will not be significantly enriched in any gene. There were 6,029 genes with a *de novo* synonymous mutation, but none of these genes was significantly enriched (enrichment p $< 2.66 \times 10^{-6}$, Bonferroni corrected for 18,762 tests; Figure 4.6c). Of note, the gene with the highest synonymous enrichment p-value from DeNovoWEST is *KAT6B* (synonymous enrichment p = 3.1×10^{-5}), which contains 9 *de novo* synonymous mutations. Six of those 9 synonymous variants are the known pathogenic synonymous variant (p.Pro1049Pro) that causes Say-Barber-Biesecker/Young-Simpson syndrome via aberrant splicing of *KAT6B* [243].

The majority (196/281; 70%) of these DeNovoWEST significant genes already had sufficient evidence of DD-association to be considered of diagnostic utility (as of late 2019) by all three centres, and I refer to them as “consensus” genes. 54/281 of these significant genes were previously considered diagnostic by one or two centres (“discordant” genes).

To discover novel DD-associated genes with greater power, I then applied DeNovoWEST only to DNMs in patients without damaging DNMs in consensus genes (I refer to this subset as ‘undiagnosed’ patients) and identified 94 significant genes (Figure 4.5c). There is a strong correlation between DeNovoWEST p-values in the full dataset compared to those in the undiagnosed-only analysis (Figure 4.6b; ρ = 0.729 for all genes with non-NA DeNovoWEST p-values in both analyses). While 61 of these genes were discordant known genes, I identified 33 putative ‘novel’ DD-associated genes. To further ensure robustness to potential mutation rate variation between genes, Kaitlin determined whether any of the putative novel DD-associated genes had significantly more synonymous variants in the Genome Aggregation

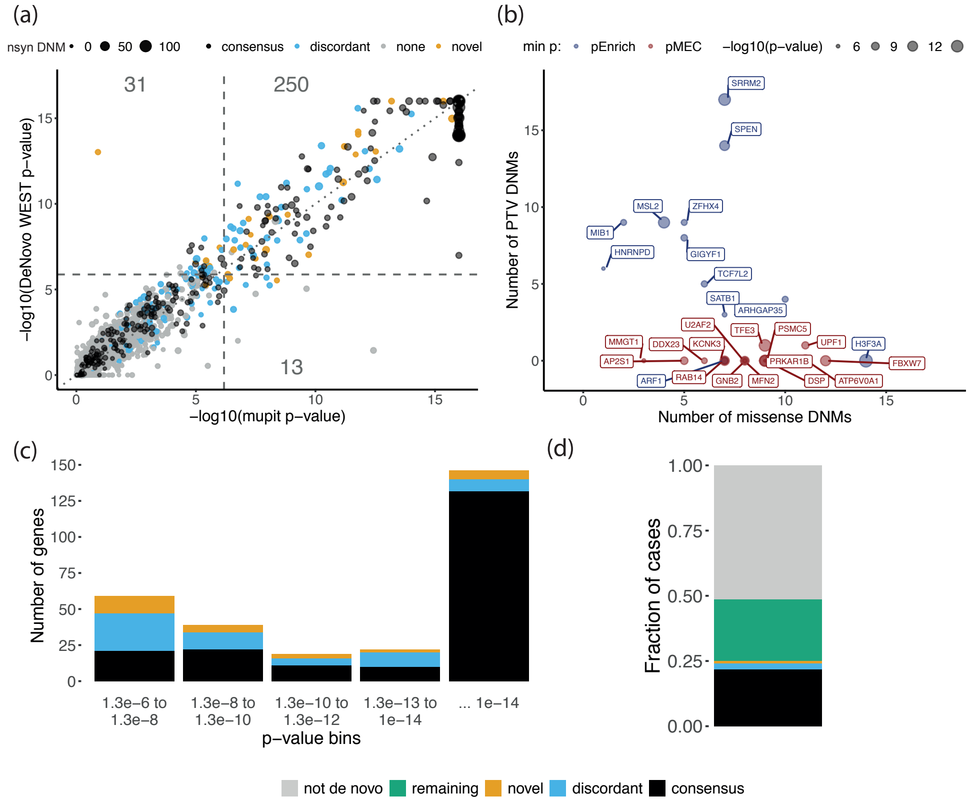


Fig. 4.5 Results of DeNovoWEST analysis. (a) Comparison of p-values generated using the new method (DeNovoWEST) versus the previous method (mupit). These are results from DeNovoWEST run on the full cohort. The dashed lines indicate the threshold for genome-wide significance. The size of the points is proportional to the number of nonsynonymous DNMs in our cohort (nsyn). The numbers describe the number of genes that fall into each quadrant. (b) The number of missense and PTV DNMs in our cohort in the 3249 novel genes. The size of the points are proportional to the -log10(p-value) from the analysis on the undiagnosed subset. The colour corresponds to which test p-value was the minimum (more significant) for these genes: pEnrich in blue, which corresponds to the overall enrichment test, or pMEC in red, which refers to the missense enrichment and clustering test. (c) The histogram depicts the distribution of p-values from the analysis on the undiagnosed subset for discordant and novel genes; p-values for consensus genes come from the full analysis. The number of genes in each bin is coloured by diagnostic gene group. (d) The fraction of cases with a nonsynonymous mutation in each diagnostic gene group. The green represents the remaining fraction of cases expected to have a pathogenic *de novo* coding mutation ("remaining") and grey is the fraction of cases that are likely to be explained by other genetic or nongenetic factors ("not *de novo*"). Figures (c) and (d) have been made by Kaitlin Samocha.

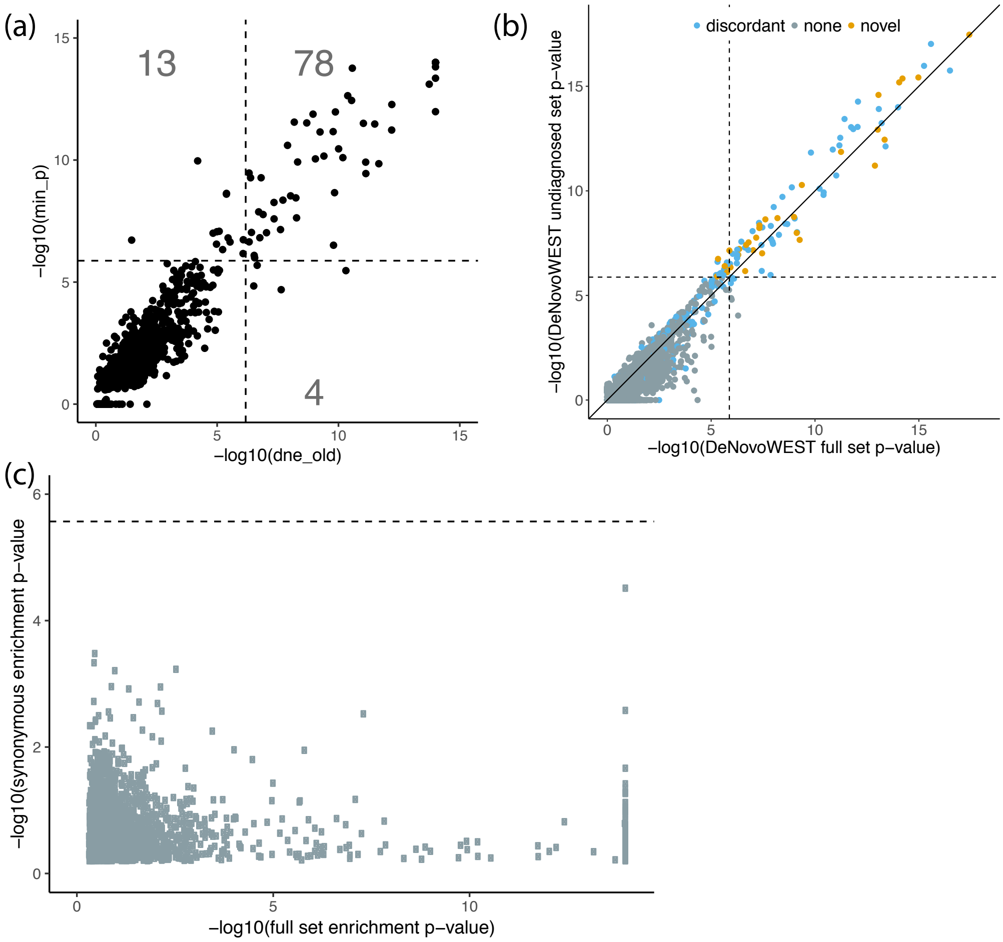


Fig. 4.6 Quality Control analyses for DeNovoWEST (a) Figure comparing p-values from published results on 4k DD trios to re-analysis of these data with the new method (DeNovoWEST). Due to constraints on the number of simulations we do not achieve p-values $< 10^{-14}$, therefore the old results are capped at this value for appropriate comparison. (b) Comparison between DeNovoWEST p-values for the full analysis vs undiagnosed-only analysis. Note that consensus genes have been removed since individuals with *de novo* nonsynonymous mutations in those genes were considered diagnosed and removed from the undiagnosed-only analysis. Genes are colored by their diagnostic list (discordant = blue; novel = orange; no list / none = gray) (c) Comparison of enrichment p-values from the full analysis vs the synonymous-only analysis. Genes with a p-value of 0 have been removed from the plot for clarity.

Database (gnomAD)[102] of population variation than expected under the null mutation model (see methods). Kaitlin identified 11/33 genes with a significant excess of synonymous variants. For these 11 genes I then repeated the DeNovoWEST test, increasing the null mutation rate by the ratio of observed to expected synonymous variants in gnomAD. Five of these genes then fell below the exome-wide significance threshold and were removed, leaving

28 novel genes, with a median of 10 nonsynonymous DNMs in our dataset (Figure 4.5b). There were 314 patients with nonsynonymous DNMs in these 28 genes (1.0% of our cohort); all DNMs in these genes were inspected in IGV[182] and, of 198 for which experimental validation was attempted, all were confirmed as DNMs in the proband. The DNMs in these novel genes were distributed approximately randomly across the three datasets (no genes with $p < 0.001$, heterogeneity test). Six of the 28 novel DD-associated genes are further corroborated by OMIM entries or publications, including *TFE3* for which patients were described in two recent publications [224, 44]. Taken together, 25.0% of individuals in the combined cohort have a nonsynonymous DNM in one of the consensus or significant DD-associated genes (Figure 4.5d).

4.3.2 Characteristics of the novel DD-associated genes and disorders

Based on an analysis from our collaborators at GeneDx of semantic similarity between Human Phenotype Ontology terms, patients with DNMs in the same novel DD-associated gene were less phenotypically similar to each other, on average, than patients with DNMs in a consensus gene ($p = 2.3\times10^{-11}$, Wilcoxon rank-sum test; Figure 4.7a) [238]. Patients with DNMs is the same novel DD-associated genes were more phenotypically similar compared to the null (pairs of random patients in the cohort) ($p = 2.0\times10^{-30}$, Wilcoxon rank-sum test). This suggests that these novel disorders less often result in distinctive and consistent clinical presentations, which may have made these disorders harder to discover via a phenotype-driven analysis or recognise by clinical presentation alone. Each of these novel disorders requires a detailed genotype-phenotype characterisation.

Overall, novel DD-associated genes encode proteins that have very similar functional and evolutionary properties to consensus genes, e.g. developmental expression patterns, network properties and biological functions (Figure 4.7b; Table 4.1). Across the properties that were considered, the only significant difference between novel and consensus genes was found in the network distance to another consensus DD gene ($p = 0.002$, Wilcoxon test). Despite the high-level functional similarity between known and novel DD-associated genes, the nonsynonymous DNMs in the more recently discovered DD-associated genes are much more likely to be missense DNMs, and less likely to be PTVs (discordant and novel; $p = 1.2\times10^{-25}$, χ^2 test). Fifteen of the 28 (54%) of the novel genes only had missense DNMs, and only a minority had more PTVs than missense DNMs. Consequently, we expect that a greater proportion of the novel genes will act via altered-function mechanisms (e.g. dominant negative or gain-of-function). For example, the novel gene *PSMC5* (DeNovoWEST $p = 2.6\times10^{-15}$) had one inframe deletion and nine missense DNMs, eight of which altered two structurally important amino acids within the 3D protein structure: p.Pro320Arg and

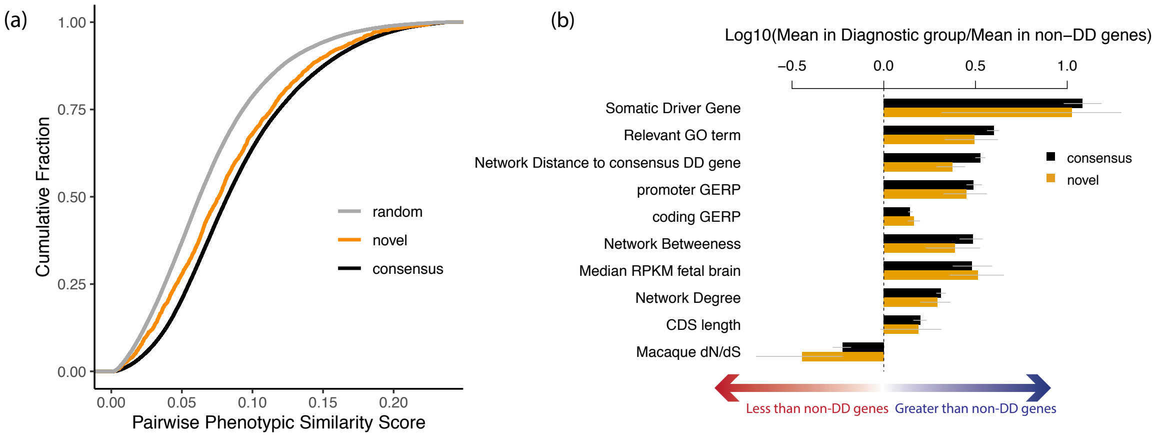


Fig. 4.7 Functional properties and mechanisms of novel genes. (a) Comparing the phenotypic similarity of patients with DNMs in novel and consensus genes. Random phenotypic similarity was calculated from random pairs of patients. Patients with DNMs in the same novel DD-associated gene were less phenotypically similar than patients with DNMs in a known DD-associated gene (p = 1.29.5 × 10^{-1338}, Wilcoxon rank-sum test). Figure made by Kevin Arvai (b) Comparison of functional properties of consensus known and novel DD genes. Properties were chosen as those known to be differential between consensus and non-DD genes

p.Arg325Trp and so is likely to operate via an altered-function mechanism. None of the novel genes exhibited significant clustering of *de novo* PTVs.

4.3.3 Recurrent mutations and potential new germline selection genes

I identified 773 recurrent DNMs (736 SNVs and 37 indels), ranging from 2-36 independent observations per DNM, which allowed me to interrogate systematically the factors driving recurrent germline mutation. I considered three potential contributory factors: (i) clinical ascertainment enriching for pathogenic mutations, (ii) greater mutability at specific sites, and (iii) positive selection conferring a proliferative advantage in the male germline, thus increasing the prevalence of sperm containing the mutation [67]. I observed strong evidence that all three factors contribute, but not necessarily mutually exclusively. Clinical ascertainment drives the observation that 65% of recurrent DNMs were in consensus genes, a 5.4-fold enrichment compared to DNMs only observed once ($p < 10^{-50}$, proportion test). Hypermutability underpins the observation that 64% of recurrent *de novo* SNVs occurred at hypermutable CpG dinucleotides[45], a 2.0-fold enrichment over DNMs only observed once ($p = 3.3 \times 10^{-68}$, χ^2 test). I also observed a striking enrichment of recurrent mutations at the haploinsufficient DD-associated gene *MECP2*, in which we observed 11 recurrently mutated

Symbol	Chr	Position	Ref	Alt	Consequence	Number recur	Likely mechanism	CpG	Somatic Driver Gene	Germline Selection Gene	DD status
PACS1	11	65978677	C	T	missense	36	activating	Yes	-	-	consensus
PPP2R5D	6	42975003	G	A	missense	22	dominant negative	-	-	-	consensus
SMAD4	18	48604676	A	G	missense	21	activating	-	Yes	-	consensus
PACS2	14	105834449	G	A	missense	13	dominant negative	Yes	-	-	discordant
MAP2K1	15	66729181	A	G	missense	11	activating	-	Yes	Yes	consensus
PPP1CB	2	28999810	C	G	missense	11	all missense/in frame	-	-	-	consensus
NAA10	X	153197863	G	A	missense	11	all missense/in frame	Yes	-	-	consensus
MECP2	X	153296777	G	A	stop gain	11	loss of function	Yes	-	-	consensus
CSNK2A1	20	472926	T	C	missense	10	activating	-	-	-	consensus
CDK13	7	40085606	A	G	missense	10	all missense/in frame	-	-	-	consensus
SHOC2	10	112724120	A	G	missense	9	activating	-	-	-	consensus
PTPN11	12	112915523	A	G	missense	9	activating	-	Yes	Yes	consensus
SMAD4	18	48604664	C	T	missense	9	activating	Yes	Yes	-	consensus
SRCAP	16	30748664	C	T	stop gain	9	dominant negative	Yes	-	-	consensus
FOXP1	3	71021817	C	T	missense	9	loss of function	Yes	-	-	consensus
CTBP1	4	1206816	G	A	missense	9	dominant negative	Yes	-	-	discordant

Table 4.2 Recurrent Mutations. *De novo* single nucleotide variants with more than 9 recurrences in the cohort annotated with relevant information, such as CpG status, whether the impacted gene is a known somatic driver or germline selection gene, and diagnostic gene group (e.g. consensus known). “Recur” refers to the number of recurrences. “Likely mechanism” refers to mechanisms attributed to this gene in the published literature

SNVs within a 500 bp window, nine of which were G to A mutations at a CpG dinucleotide. *MECP2* exhibits a highly significant twofold excess of synonymous mutations within the Genome Aggregation Database (gnomAD) population variation resource[102], suggesting that locus-specific hypermutability might explain this observation.

To assess the contribution of germline selection to recurrent DNMs, I initially focused on the 12 known germline selection genes (*FGFR2, FGFR3, PTPN11, HRAS, KRAS, RET, BRAF, CBL, MAP2K1, MAP2K2, RAF1, SOS1*), which all operate through activation of the RAS-MAPK signalling pathway[136, 137]. I identified 39 recurrent DNMs in 11 of these genes, 38 of which are missense. To determine if the observed mutations in germline selection genes are known to be activating, I first confirmed that these were all genes known to be acting through gain-of-function mechanisms. All of these genes are known monoallelic DD-associated genes annotated as having activating mutation consequences according to DDG2P[236]. I then confirmed that all these recurrent mutations are listed as pathogenic or likely pathogenic variants in ClinVar[115]. As expected, given that hypermutability is not the driving factor for recurrent mutation in these germline selection genes, these 39 recurrent DNMs were depleted for CpGs relative to other recurrent mutations (6/39 vs 425/692, p = 3.4×10^{-8}, χ^2 test). Positive germline selection has been shown to be capable of increasing the apparent mutation rate more strongly than either clinical ascertainment (10-100× in this dataset) or hypermutability (~10× for CpGs)[67]. However, only a minority of the most highly recurrent mutations in this dataset are in genes that have been previously associated with germline selection. Nonetheless, several lines of evidence suggested that the majority of these most highly recurrent mutations are likely to confer a germline selective advantage. Based on the recurrent DNMs in known germline selection genes, DNMs under germline selection should be more likely to be activating missense mutations, and should be less enriched for CpG dinucleotides. Table 4.2 shows the 16 *de novo* SNVs observed nine or more times in our DNM dataset, only two of which are in known germline selection genes (*MAP2K1* and *PTPN11*). All but two of these 16 *de novo* SNVs cause missense changes, all but two of these genes cause disease by an altered-function mechanism, and these DNMs were depleted for CpGs relative to all recurrent mutations. Two of the genes with highly recurrent *de novo* SNVs, *SHOC2* and *PPP1CB*, encode interacting proteins that are known to play a role in regulating the RAS-MAPK pathway, and pathogenic variants in these genes are associated with a Noonan-like syndrome[244]. Moreover, two of these recurrent DNMs are in the same gene *SMAD4*, which encodes a key component of the TGF-beta signalling pathway, potentially expanding the pathophysiology of germline selection beyond the RAS-MAPK pathway. Confirming germline selection of these mutations will require deep sequencing of testes and/or sperm[137].

4.3.4 Evidence for incomplete penetrance and pre/perinatal death

Nonsynonymous DNMs in consensus or significant DD-associated genes accounted for half of the exome-wide nonsynonymous DNM burden associated with DD (Figure 4.5d). Despite the identification of 285 significantly DD-associated genes, there remains a substantial burden of both missense and protein-truncating DNMs in unassociated genes (those that are neither significant in my analysis nor on the consensus gene list). The remaining burden of protein-truncating DNMs is greatest in genes that are intolerant of PTVs in the general population (Figure 4.8a). Similarly, while I observed that unassociated genes were, overall, significantly less likely to be expressed in the fetal brain ($p = 4.42 \times 10^{-30}$, proportion test), I found that within genes intolerant of PTVs (pLI>0.9), unassociated genes were just as likely as significant genes to be expressed in the fetal brain ($p = 0.09$, proportion test; Figure 4.9). This suggests that more haploinsufficient (HI) disorders await discovery. The PTV enrichment dropped even further after removing nominally significant genes in my analysis which suggests that a larger sample size will increase the power to detect these (Figure 4.8b). I observed that PTV mutability (estimated from a null germline mutation model) was significantly lower in unassociated genes compared to DD-associated genes (p $=4.5 \times 10^{-68}$, Wilcox rank-sum test 4.10a), which leads to reduced statistical power to detect

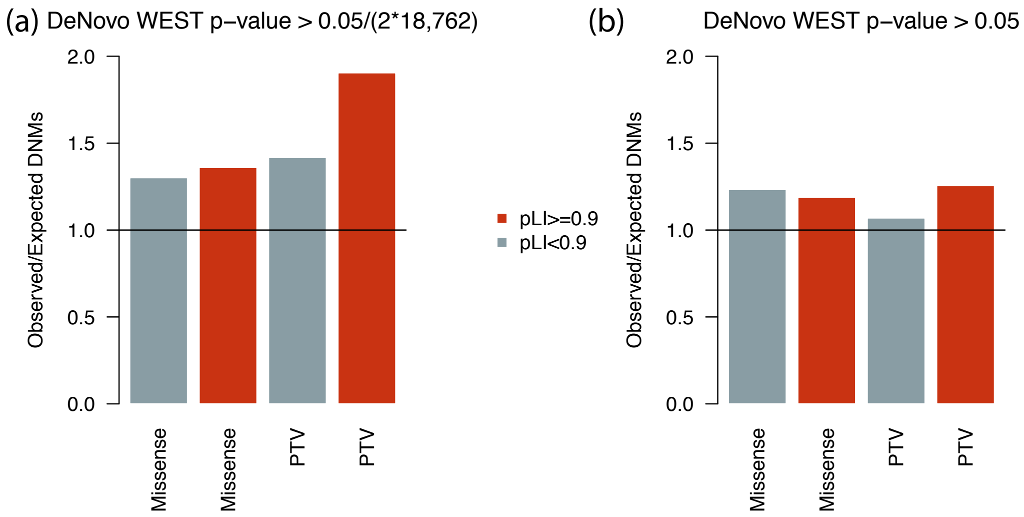


Fig. 4.8 DNM enrichment in non-significant genes (a) The enrichment of missense and PTVs separated by high pLI (red) and low pLI (blue). Unassociated genes are defined as those not on any diagnostic list (not consensus or discordant) and not significant in our analysis. (b) The same as (a) except here the genes have been subsetted further to those that are not nominally significant in our analysis (unadjusted p-value >0.05)

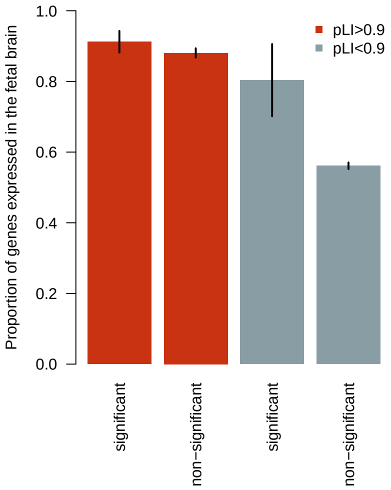


Fig. 4.9 Comparison of proportion of genes expressed in fetal brain. The proportion of genes between significant and non-significant genes in our analysis split by low (<0.9) and high (>0.9) pLI. Significant genes also includes all genes on the consensus diagnostic list.

DNM enrichment in unassociated genes. This is consistent with the hypothesis that many more HI disorders await discovery.

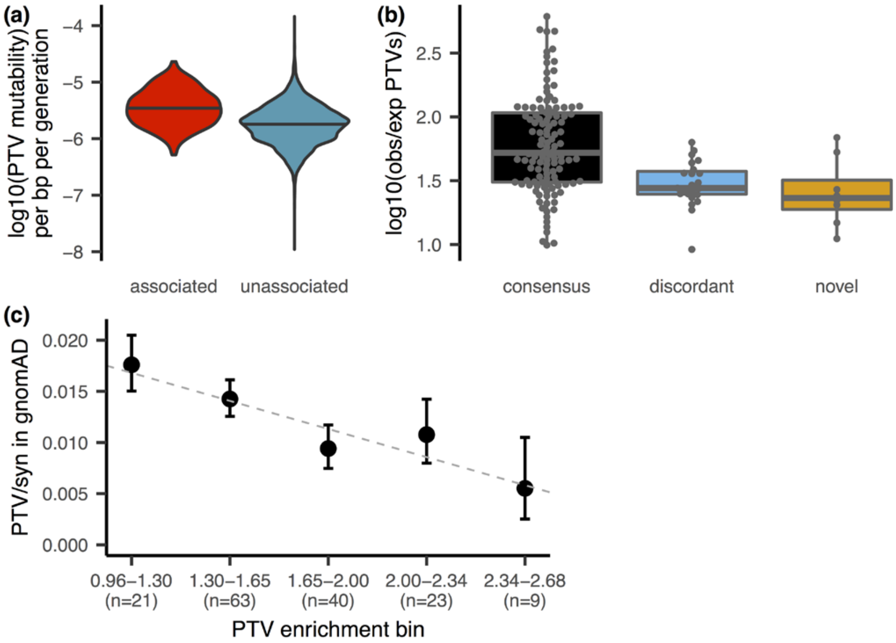


Fig. 4.10 Impact of penetrance on power. (a) PTV mutability is significantly lower in genes that are not significantly associated to DD in the analysis ("unassociated", coloured blue) than in DD-associated genes ("associated", coloured red; p = 4.5×10^{-68}, Wilcox rank sum test). (b) Distribution of PTV enrichment in significant, likely haploinsufficient, genes by diagnostic group. (c) Comparison of the PTV enrichment in the cohort vs the PTV to synonymous ratio found in gnomAD, for genes that are significantly enriched for the number of PTV mutations in the cohort (without any variant weighting). PTV enrichment is shown as log10(enrichment). There is a significant negative relationship (p = 0.031, weighted regression). Figure (c) courtesy of Kaitlin Samocha

A key parameter in estimating statistical power to detect novel HI disorders is the fold-enrichment of *de novo* PTVs expected in as yet undiscovered HI disorders. I observed that novel DD-associated HI genes had significantly lower PTV enrichment compared to the consensus HI genes (p = 0.005, Wilcox rank-sum test; 4.10b). Two additional factors that could lower DNM enrichment, and thus power to detect a novel DD-association, are reduced penetrance and increased pre/perinatal death, which here covers spontaneous fetal loss, termination of pregnancy for fetal anomaly, stillbirth, and early neonatal death. To evaluate incomplete penetrance, Kaitlin investigated whether HI genes with a lower enrichment of protein-truncating DNMs in our cohort are associated with greater prevalences of PTVs in the general population. She observed a significant negative correlation (p = 0.031, weighted linear regression) between gene-specific PTV enrichment in our cohort and the gene-specific ratio of PTV to synonymous variants in the gnomAD dataset of population variation[102], suggesting that incomplete penetrance does lower *de novo* PTV enrichment in individual genes in the cohort (Figure 4.10c).

A structural malformation (detected via ultrasound) during pregnancy is associated with pre/perinatal death, which here encompasses spontaneous fetal loss, stillbirth, early neonatal death, and termination of pregnancy for fetal anomaly. To understand the impact that pre/perinatal death may have had on our power to detect DD-associated genes a clinician (Allison Yeung) assigned each consensus DD gene known to be haploinsufficient (n=217) a low, medium or high likelihood of a patient with a pathogenic mutation in the gene presenting with a structural malformation on ultrasound. To verify that this classification was valid, I compared the proportion of individuals in the DDD study with nonsynonymous mutations in the three gene groups that were reported to have an abnormal scan during pregnancy (Figure 4.11a). I found that the proportion of patients with an abnormal ultrasound for those with a nonsynonymous DNM in a gene with a high or medium likelihood of abnormal ultrasound was significantly higher than for those with a DNM in a gene with a low likelihood classification (12.8% (low) vs 28.0% (medium/high), χ^2 p = 7.24×10^{13}). I also looked at DNMs called in 640 trios from the Prenatal Assessment of Genomes and Exomes (PAGE) study and found that there was a higher enrichment of non synonymous DNMs in genes with a medium/high likelihood of presenting with an ultrasound abnormality compared to the genes with a low likelihood but this was not significant (2.3 (low) vs 4.8 (medium/high) enrichment, p = 0.052, Poisson test; Figure 4.11d) [133].

Having verified this classification, I then calculated the ratio of the number of observed *de novo* PTVs in each group to the total number of expected *de novo* PTVs across each group. I observed that the fold-enrichment of protein-truncating DNMs in consensus HI DD-associated genes in the cohort was significantly lower for genes with a medium or high likelihood of presenting with a prenatal structural malformation (p = 4.6×10^{-5}, Poisson test, Figure 4.11c), suggesting that pre/perinatal death decreases power to detect some novel DD-associated disorders. For the DDD data, I also regressed the proportion of individuals with a nonsynonymous mutation in a consensus HI gene who had an abnormal ultrasound against the observed PTV enrichment in that gene but did not find a significant regression coefficient for proportion of abnormal ultrasounds (p = 0.33, quasipoisson GLM; Figure 4.11b).

To assess whether mutations in novel genes are more likely to be associated with pre/perinatal death than consensus genes I compared the nonsynonymous *de novo* enrichment of these two groups in PAGE. I did not find a significant difference between the enrichment in novel genes (2.9) compared to consensus genes (3.4) (p = 0.44, Poisson test) however this analysis is not well powered. I also did not find that individuals in DDD with mutations

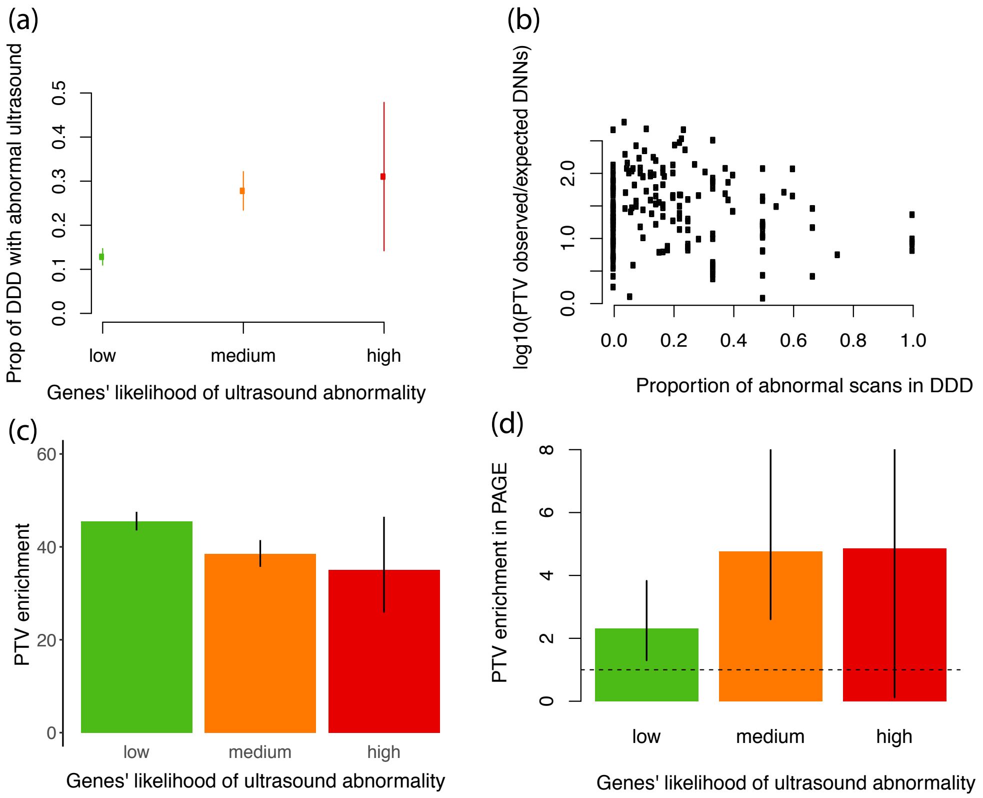


Fig. 4.11 Investigating impact of pre/perinatal death on power (a) The proportion of DDD probands that have been reported as presenting with an abnormal ultrasound furing pregnancy that have nonsynonymous DNMs in consensus genes. This has been split by whether the gene falls in the low, medium of high category of the clinician classified likelihood of presenting with a ultrasound abnormality. This has been coloured by red (high), orange (medium) and low (green). (b) This shows observed PTV enrichment in the cohort for each consensus known haploinsufficient gene against the proportion of individuals with a nonsynoymous DNM in those genes that have been reported in DDD as having an abnormal scan (c) Overall *de novo* PTV enrichment (observed / expected PTVs) across genes grouped by their clinician-assigned likelihood of presenting with a structural malformation on ultrasound during pregnancy in the cohort. PTV enrichment is significantly lower for genes with a medium or high likelihood compared to genes with a low likelihood (p = 4.5×10^{-5}, Poisson test) (d) Overall *de novo* PTV enrichment (observed / expected PTVs) across genes grouped by their clinician-assigned likelihood of presenting with a structural malformation on ultrasound during pregnancy in the PAGE cohort.

in novel genes were significantly more likely to have an abnormal ultrasound compared to consensus genes (proportion abnormal for novel 0.24, proportion abnormal for consensus 0.17; χ^2 test p = 0.08) but again there is not sufficient power here.

4.3.5 Modelling reveals hundreds of DD genes remain to be discovered

To understand the likely trajectory of future DD discovery efforts, I downsampled the current cohort (to 5k, 10k, 15k, 20k and 25k individuals) and reran the enrichment analysis (Figure 4.12a). I observed that the number of significant genes has not yet plateaued. Increasing sample sizes should result in the discovery of many novel DD-associated genes. To estimate how many haploinsufficient genes might await discovery, I modelled the likelihood of the observed distribution of protein-truncating DNMs among genes as a function of varying numbers of undiscovered HI DD genes and fold-enrichments of protein-truncating DNMs in those genes. I found that the remaining HI burden is most likely spread across ~1000 genes with ~10-fold PTV enrichment (Figure 4.12b). This fold enrichment is three times lower than in known HI DD-associated genes, suggesting that incomplete penetrance and/or pre/perinatal death is much more prevalent among undiscovered HI genes. I modelled the missense DNM burden separately which allowed for different distribution of missense enrichments, as modelled by the gamma distribution, across DD genes. These distributions represent different scenarios for the proportion of genes acting via altered-function or loss-of-function mechanisms. I observed that the most likely architecture of undiscovered DD-associated genes is one that comprises over 1000 genes with a substantially lower fold-enrichment (~3 fold) than in currently known DD-associated genes (Figure 4.13 b). The most likely missense

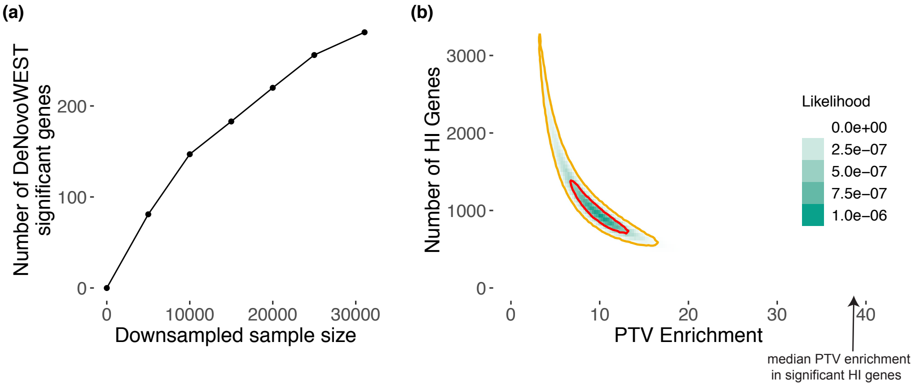


Fig. 4.12 Exploring the remaining number of DD genes. (a) Number of significant genes from downsampling full cohort and running DeNovoWEST's enrichment test. (b) Results from modelling the likelihood of the observed distribution of *de novo* PTV mutations. This model varies the numbers of remaining haploinsufficient (HI) DD genes and PTV enrichment in those remaining genes. The 50% credible interval is shown in red and the 90% credible interval is shown in orange. Note that the median PTV enrichment in significant HI genes (shown with an arrow) is 39.7

enrichment distribution reflected the scenario where most of the missense mutations were acting via loss-of-function mechanisms (γ shape parameter of 20).

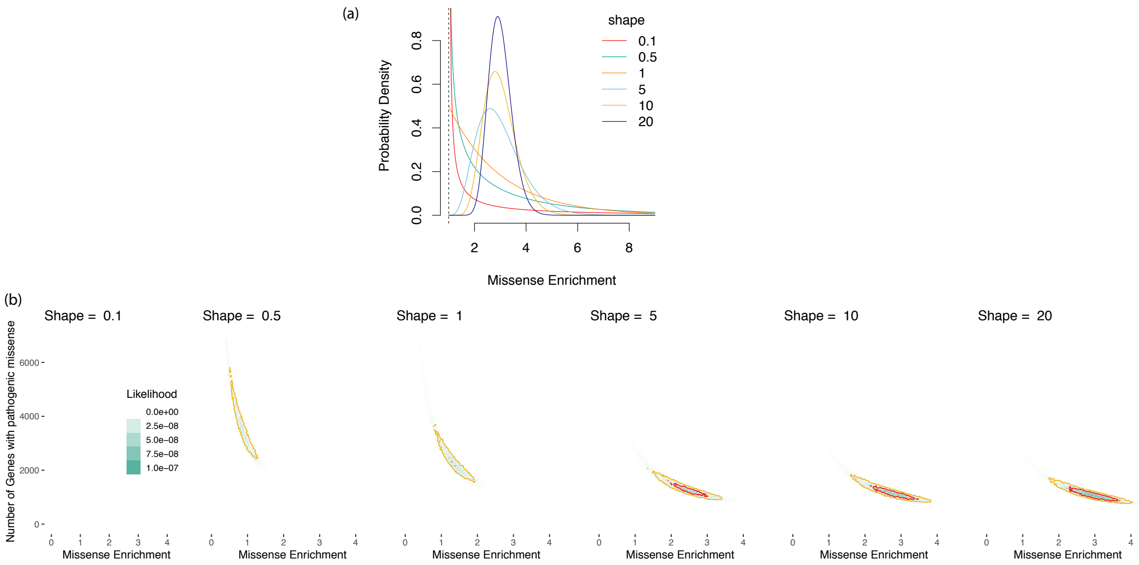


Fig. 4.13 Likelihood model for missense DNM enrichment(a) Depiction of γ distribution for 6 shape values used in simulations. Here the mean of each distribution is set at 2. (b) Likelihood of scenario under variety of shapes for γ distribution and under varying values of missense enrichment and number of genes with pathogenic missense variants. The 90% (yellow line) and 50% (red line) credible intervals are calculated across all shapes considered.

A sample size of ~350,000 parent-offspring trios would be needed to have 80% power to detect a 10-fold enrichment of protein-truncating DNMs for a gene with the median PTV mutation rate among currently unassociated genes (assuming Poisson test on PTV count enrichment). Using this inferred 10-fold enrichment among undiscovered HI genes, from our current data I can evaluate the likelihood that any gene in the genome is an undiscovered HI gene, by comparing the likelihood of the number of *de novo* PTVs observed in each gene to have arisen from the null mutation rate or from a 10-fold increased PTV rate. Among the ~19,000 non-DD-associated genes, ~1,200 were more than three times more likely to have arisen from a 10-fold increased PTV rate, whereas ~7,000 were three times more likely to have no *de novo* PTV enrichment.

4.4 Discussion

In this chapter, I identified 28 novel developmental disorders by developing an improved statistical test for mutation enrichment and applying it to a dataset of exome sequences from 31,058 children with developmental disorders, and their parents. These 28 novel genes account for up to 1.0% of the cohort, and inclusion of these genes in diagnostic workflows will catalyse increased diagnosis of similar patients globally. The value of this study for improving diagnostic yield extends well beyond these 28 novel genes; once newly validated discordant genes are included, the total number of genes added to the diagnostic workflows of the three participating centres ranged from 48-65 genes. I have shown that both incomplete penetrance and pre/perinatal death reduce our power to detect novel DDs postnatally, and that one or both of these factors are likely operating considerably more strongly among undiscovered DD-associated genes. In addition, I have identified a set of highly recurrent mutations that are strong candidates for novel germline selection mutations, which would be expected to result in a higher than expected disease incidence that increases dramatically with increased paternal age. This study represents the largest collection of DNMs for any disease area, and is approximately three times larger than a recent meta-analysis of DNMs from a collection of individuals with autism spectrum disorder, intellectual disability, and/or a developmental disorder[31]. The analysis included DNMs from 24,348 previously unpublished trios, and I identified ~2.3 times as many significantly DD-associated genes as this previous study when using Bonferroni-corrected exome-wide significance (285 vs 124). In contrast to meta-analyses of published DNMs, the harmonised filtering of candidate DNMs across cohorts in this study should protect against results being confounded by substantial cohort-specific differences in the sensitivity and specificity of detecting DNMs.

Here I inferred indirectly that developmental disorders with higher rates of detectable prenatal structural abnormalities had greater pre/perinatal death. The potential size of this effect can be quantified from the recently published PAGE study of genetic diagnoses in a cohort of fetal structural abnormalities[133]. In this latter study, genetic diagnoses were not returned to participants during the pregnancy, and so the genetic diagnostic information itself could not influence pre/perinatal death. In the PAGE study data, 69% of fetuses with a genetically diagnosable cause for this anomaly died perinatally or neonatally, with termination of pregnancy, fetal demise and neonatal death all contributing. This emphasises the substantial impact that pre/perinatal death can have on reducing the ability to discover novel DDs from postnatal recruitment alone, and motivates the integration of genetic data from prenatal, neonatal and postnatal studies in future analyses.

To empower mutation enrichment testing, I estimated positive predictive values (PPV) of each DNM being pathogenic on the basis of their predicted protein consequence, CADD score, selective constraint against heterozygous PTVs in the gene (s_{het}), and, for missense variants, presence in a region under missense constraint in the general population [107, 24, 186]. These PPVs should also be highly informative for variant prioritisation in the diagnosis of dominant developmental disorders. Further work is needed to see whether these PPVs might be informative for recessive developmental disorders, and in other types of dominant disorders. More generally, empirically-estimated PPVs based on variant enrichment in large datasets may be similarly informative in many other disease areas.

The approach taken here is statistically conservative in identifying DD-associated genes. In two previous published studies from the DDD, using the same significance threshold, 26 novel DD-associated genes were identified[215, 41]. All 26 are now regarded as being diagnostic, and have entered routine clinical diagnostic practice. There are 184 consensus genes that did not cross the significance threshold in this study. It is likely that many of these cause disorders that were under-represented in this study due to the ease of clinical diagnosis on the basis of distinctive clinical features or targeted diagnostic testing. These ascertainment biases are, however, not likely to impact the representation of novel DDs in this cohort. The modelling also suggested that likely over 1,000 DD-associated genes remain to be discovered, and that reduced penetrance and pre/perinatal death will reduce power to identify these genes through DNM enrichment. Identifying these genes will require both improved analytical methods and greater sample sizes. As sample sizes increase, accurate modelling of gene-specific mutation rates becomes more important. In this analyses of 31,058 trios, there was evidence that mutation rate heterogeneity among genes can lead to over-estimating the statistical significance of mutation enrichment based on an exome-wide mutation model. An important future direction would be to develop more granular mutation rate models, based on large-scale population variation resources, to ensure that larger studies are robust to mutation rate heterogeneity.

The variant-level weights used by DeNovoWEST could be improved over time. As reference population samples, such as gnomAD, increase in size, weights based on selective constraint metrics (e.g. s_{het}, regional missense constraint) will improve. Weights could also incorporate more functional information, such as expression in disease-relevant tissues [102]. For example, I observed that DD-associated genes are significantly more likely to be expressed in fetal brain (Figure 4.9). Furthermore, novel metrics based on gene co-regulation networks can predict whether genes function within a disease-relevant pathway[42]. As a cautionary note, including more functional information may increase power to detect some novel disorders while decreasing power for disorders with pathophysiology different from known disorders. Variant-level weights could be further improved by incorporating other variant prioritisation metrics, such as upweighting variants predicted to impact splicing, variants in particular protein domains, or variants that are somatic driver mutations during tumorigenesis. In developing DeNovoWEST, I initially explored applying both variant-level weights and gene-level hypothesis weights in separate stages of the analysis, however, subtle but pervasive correlations between gene-level metrics (e.g. s_{het}) and variant-level metrics (e.g. regional missense constraint, CADD) presents statistical challenges to implementation. Finally, the discovery of less penetrant disorders can be empowered by analytical methodologies that integrate both DNMs and rare inherited variants, such as TADA [78]. Nonetheless, using current methods, I estimated that ~350,000 parent-child trios would need to be analysed to have ~80% power to detect HI genes with a 10-fold PTV enrichment. Discovering non-HI disorders will need even larger sample sizes. Reaching this number of sequenced families will be impossible for an individual research study or clinical centre, therefore it is essential that genetic data generated as part of routine diagnostic practice are shared with the research community such that it can be aggregated to drive discovery of novel disorders and improve diagnostic practice.

Chapter 5

Discussion

5.1 Summary of Findings

Understanding variation and consequences of germline mutation is paramount for understanding evolutionary process, biological mechanisms of mutation and causes of genetic disease. In this thesis I have described three distinct projects that have attempted to explore different aspects of germline mutation.

In Chapter 2 I examined the mutational origins and pathogenic consequences of MNVs, an important source of genetic variation. I found that the most frequent type of MNV was at adjacent nucleotides and that more than half of these were likely due to a single mutational event. I then confirmed observations from previous studies that there appear to be several mutational processes creating different types of MNVs. I estimated the MNV mutation rate to be 1.6% that of SNVs. Most population genetics models assume that mutations arise from independent events and ignoring clustered mutations can lead to incorrect inferences about positive selection. Understanding the mutational spectra and mutation rate of MNVs will help inform future refinements to these models. The presence of a possible MNV mutator phenotype that I observed in individuals in the DDD study may further complicate these refinements. I found that MNVs in protein coding sequences are on average more pathogenic than SNVs. Even when the MNV falls within a single codon, they can create a larger physiochemical change and have a greater functional impact than an SNV in the same codon. I identified 10 pathogenic *de novo* MNVs within the DDD study. My findings demonstrate that MNVs constitute a unique class of variant in both mutational origin and functional impact. This has implications in how variants are annotated in the future and correct annotation is important in furthering our understanding of their role in evolution and disease.

In Chapter 3 I describe my work on the identification and characterisation of germline hypermutators. I identified fifteen individuals, 1 from the Deciphering Developmental Disorders (DDD) study and 14 from the 100,000 Genomes Project (100kGP), with a significantly increased number of *de novo* SNV mutations (DNMs) that ranged from a 2 to 7 fold enrichment compared to the expected number. The DNMs in these individuals exhibited distinctive mutational spectra, some of which mapped on to known somatic mutational signatures identified in cancers as well currently unknown mutational signatures. I found that for the majority of these individuals the excess mutations were paternally derived implicating the father as a possible germline hypermutator. In two of these fathers I identified rare nonsynonymous homozygous variants in genes known to be associated with DNA repair, *MPG* and *XPC*. *MPG* is known to be involved in the base excision repair pathway while *XPC* is associated with the nucleotide excision repair pathway. The variant in *XPC* is a pathogenic PTV known to cause xeroderma pigmentosum with which the father has previously been diagnosed. This finding suggests that other *XPC* carriers may have a similar mutator phenotype and should be followed up with additional pedigree WGS studies. Germline hypermutation accounted for 7% of variation in the germline mutation rate in 100kGP despite only affecting 14 individuals. I estimated that parental age accounted for ~70% of germline mutation rate variation which leaves ~20% of variance unaccounted for. These findings suggest that defects in DNA repair genes can dramatically increase the germline mutation rate. However, I found that the presence of a nonsynonymous variant across different subsets of DNA repair genes did not significantly impact the number of DNMs per person across the cohort, even those variants known to increase the somatic mutation rate. This suggest that at most a subset of variants in DNA repair genes associated with cancer also affect the germline and possibly that hypermutation is more often a recessive phenotype and heterozygous variants may not increase mutation rates. This is also demonstrated by my findings on the impact of germline PTVs in the gene *MBD4*. I found that these variants, that are known to be associated with a somatic mutator phenotype, have no detectable effect on germline mutation rate. These analyses have provided new insights into how genetic variation can impact individual germline mutation rate and that hypermutation accounts for a substantial fraction of variance in germline mutation rates. The remaining unexplained variance calls attention to the need to investigate additional sources of variation such as polygenic or environmental contributions. This will likely require being able to cheaply and accurately assay germline mutation rates to be able to perform a genome-wide association study.

My final project, described in Chapter 4, shifts away from investigating variation in the types and rates of DNMs to their role in causing rare disease. I integrated health care and research exome sequences and analysed *de novo* mutations in a cohort of 31,058 parent offspring trios with developmental disorders. I developed an improved statistical framework that increased power to identify gene-specific enrichments. I applied this on the cohort and identified 285 genes significantly associated with DD, including 28 that have not previously been robustly associated with DDs. Despite detecting more DD-associated genes, ~50% of the excess of DNMs in protein-coding genes remained unaccounted for. I performed a down-sampling analysis and found that the discovery of DD-associated genes has not plateaued and that increasing samples sizes should result in discovery of many novel DD-associated genes. I modelled the likelihood of the observed distribution of both missense and protein-truncating DNMs and my results suggest that over 1,000 novel DD-associated genes await discovery. I found that the undiscovered genes are likely to be less penetrant than the currently known genes and that pre/peri-natal death is reducing power to detect novel DD-associated genes. This has important implications for how the field approaches the discovery of the remaining genes and that this will require improved analytical methods as well as increased sample sizes. A substantial role for incomplete penetrance suggests that combining inherited and *de novo* variation may aid discovery of novel DD-associated genes. I also identified a set of highly recurrent mutations which are strong candidates for novel germline selection mutations. These need to be examined further as they may result in a much higher disease incidence than expected which would increase with paternal age.

5.2 Limitations and future directions

There are several limitations in the work I have described in this thesis and these can help highlight future avenues of research to further our understanding of germline mutation and its role in rare disease.

This work highlights the deficiencies of current mutational models in addressing mutation heterogeneity and different types of genetic variation. In Chapter 4, I found that some of the initially significant DD-associated genes were enriched for synonymous variants in gnomAD and this could be evidence of mutation rate heterogeneity across genes. To improve the mutational model one can start to include additional annotation of features known to influence mutation rate. For example, recently developed models have started to incorporate methylation status which impacts CpG mutation rates[102, 22]. Other factors known to influence mutation rate such as GC content, proximity to recombination hotspots or replication timing could also be incorporated. Somatic mutational models have started to include these and a plethora of other local genomic features ([13, 120, 171, 141]). However these models have the benefit of being able to be trained on a large number of somatic mutations in normal tissue which is not currently tractable with germline mutations. The average number of DNMs per person is ~70 and require whole genome sequencing of both the child and two parents. Currently most large collections of DNMs are in disease cohorts, such as for DD or autism, where DNMs are enriched and due to clinical ascertainment the mutations would not be distributed as expected under a null germline mutational model. A large dataset of *de novo* mutations in healthy individuals would be useful in training a null germline mutational model. These DNMs would also help to inform how we can include MNVs in a null mutational model and help us further investigate individual mutation rate variation. DNMs in healthy individuals would represent a combination of both mutation and negative selection, to examine mutation alone sequencing of gametes is needed however this is very difficult to amass. Trio sequencing of a healthy population would also take time to build and a more feasible approach currently would be to train a mutational model on rare variation in large healthy population cohorts as an approximation for germline mutation.

In all three projects I focused primarily on SNVs and expanding these to include indels would be informative. MNVs which include multiple indels, or possibly include an SNV together with an indel, are likely to have a larger functional impact compared to multiple SNVs and it would be interesting to explore if there are specific mutational mechanisms that may create these. When investigating germline hypermutation in Chapter 3, there was some evidence that some individuals have an excess of *de novo* indels despite having an expected number of SNVs. Since the indel calling was less sensitive than SNVs these candidate *de novo* indels would need to be carefully examined to ensure they are real. In Chapter 4, I was not able to estimate the positive predictive value of being pathogenic directly for indels and I used weights from missense and nonsense variants as proxies for inframe and frameshift mutations respectively. This was due to the fact that I did not have a mutational model for indels that could reliably calculate the expected number of these mutation types within the exome. Developing a mutational model for indels would be an important development in the field. Indels are inherently more complex to create a mutational model than SNVs as there are added dimensions of the length of the indel as well as whether an insertion/deletion is created. Training such a model requires a large dataset of high quality indels which is difficult to compile since indel calling tends to be less accurate than SNVs and indels have a much lower mutation rate compared to SNVs.

Sample size is a limitation in all three of the projects I describe and building larger parent-offspring trio sequencing datasets will be key in addressing further unanswered questions from this thesis. In the context of developmental disorders, despite pooling over 31,000 exome-sequenced trios from two research and one healthcare generated datasets in Chapter 4, I found that larger sample sizes of 100,000 trios would be needed to detect additional genes associated with developmental disorders. Achieving sample sizes of this size is only possible through collaboration with healthcare generated data. GeneDx continues to grow their dataset through clinical genomics testing. In the UK the creation of the NHS Genomics Medicine service means that children with rare diseases, as well as certain rare disease and cancer in adults, will be eligible for whole genome sequencing. The 100,000 Genomes Project has now been expanded and the next aim is to sequence 5 million genomes via research and industry partners with the NHS and UK biobank contributing to 1 million of these genomes. These datasets will include a large number of parent-offspring trios and, as well as helping our study of rare genetic disease, could also be useful in understanding properties of germline mutation. Larger sample sizes are needed to untangle the impact of genetic variation on germline mutation rate. These efforts could focus on a curated set of variants in genes already known to be involved in DNA repair or be more agnostic and look across all genes. These studies could also help us to understand the differences between variants that impact somatic and germline mutation differently. Inclusion of families with more than one child or even large families in these datasets will also allow us to further investigate how germline mutation rate can vary between, and within, families.

A limitation to much of my work in this thesis is that it has been focussed on coding variation. The increasing sample sizes of WGS datasets will allow more interrogation of the non-coding genome. In Chapter 4 the study of DNMs in DD was limited to coding regions of the genome. Previous work conducted in the Hurles group started to address the role of DNMs in highly conserved fetal brain regulatory elements, however it was concluded that much larger sample sizes would be needed to have the power to identify specific elements associated with DD[201]. Short et al. also highlighted the the importance of improved pathogenicity annotations in the non-coding genome. There has been several methods developed recently to predict pathogenicity by combining many different genomic features that prioritise non-coding variants[107, 230, 85, 246, 203, 199, 88]. As WGS datasets increase in size and there are a sufficient amount of DNMs, these annotations could be included in the DeNovoWEST framework by incorporating them into the weights. Improvements in pathogenicity annotation will also be key in furthering other work from this thesis. When searching for possible germline mutator variants in hypermutated individuals in Chapter 3 I was also restricted to coding variants but some of these variants of interest may lie in regulatory regions. My study of MNVs in Chapter 2 was restricted to the exome and even with large WGS we do not currently have the power to detect the functional impact of non-coding MNVs. Better annotation of regulatory regions may allow us to see differing levels of enrichment in disease cohorts of MNVs compared to SNVs.

In addition to large scale whole genome sequencing of families, single-cell sequencing of sperm will be an important tool to investigate different properties of male germline mutation. As mentioned in Chapter 3, current estimates of individual germline mutation rate are restricted to averaging the number of DNMs across offspring however this is often based on a single observation and is a combination of the maternal and paternal germline mutation rates. The development of single-cell sequencing and its application to sperm will allow more accurate estimation of individual mutation rates. These technologies should also allow examination of within individual mutation rates as well as clonal dynamics occurring in the testes and could help identify further mutations that undergo germline selection. In the context of germline hypermutation, sequencing of sperm would confirm if hypermutation occurs in all sperm and explore whether some mutator variants might be mosaic, specific to the germline and thus undetectable in soma-derived DNA.

My thesis was focused on germline mutation and so in Chapter 4 I only assessed the contribution of DNMs to developmental disorders. However, these analyses suggested that reduced penetrance of mutations was impacting the power to detect novel genes associated with DD and that incorporating inherited variation together with DNMs will be an important next step in identifying these genes. Pre/peri-natal lethality also reduces power to detect DD-associated genes and expanding pre-natal fetal sequencing cohorts may be helpful in discovering these genes.

5.3 Concluding remarks

The advent of exome and whole genome sequencing has allowed for close examination of germline mutation in the last decade. Direct identification of DNMs from families has led to improved estimation of mutation rates and has provided important insights into factors that can influence this rate at a genome, individual and population level. These advances have enabled the construction of a mutational model which has aided identification of genes associated with rare genetic disease and led to countless diagnoses for families which may pave the way for possible treatments. However we have only started to scratch the surface of understanding germline mutation. The increase in sample size of sequenced cohorts in the last few years has revealed additional complexities in sources of variation such as those discussed in this thesis; from the impact of non-independent mutations to the existence of hypermutators. In the future, more detailed characterisation of individual environment and geospatial analyses of mutation rates may also uncover the impact of environmental mutagens. The increase in sample size has also started to uncover the deficiencies in current mutational models which can confound our detection of disease associations. Incorporation of what we have learnt about the process of germline mutation into improved mutational models is paramount, both for evolutionary studies and to improve how we assess the contribution of

DNMs to disease. Even with improvements to the mutational model, integration of data from vast numbers of families will be needed to achieve a more complete picture of the genes involved in rare disease. Global collaboration between healthcare systems and research initiatives will be key in amassing the amounts of data needed and this will further not only our understanding of rare disease but also of germline mutation as a whole.

References

[1] Acuna-Hidalgo, R., Veltman, J. A., and Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. *Genome Biology*, 17(1):241.

[2] Adewoye, A. B., Lindsay, S. J., Dubrova, Y. E., and Hurles, M. E. (2015). The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline. *Nature Communications*, 6(1):1–8.

[3] Aggarwala, V. and Voight, B. F. (2016). An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics*, 48(4):349.

[4] Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Ng, A. W. T., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., et al. (2020). The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101.

[5] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463):415.

[6] Allen, A. S., Berkovic, S. F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E. E., Epstein, M. P., Glauser, T., Goldstein, D. B., Han, Y., et al. (2013). De novo mutations in epileptic encephalopathies. *Nature*, 501(7466):217–221.

[7] Amirnovin, R. (1997). An analysis of the metabolic theory of the origin of the genetic code. *Journal of Molecular Evolution*, 44(5):473–476.

[8] Amos, W. (2010). Even small snp clusters are non-randomly distributed: is this evidence of mutational non-independence? *Proceedings of the Royal Society B: Biological Sciences*, 277(1686):1443–1449.

[9] Amos, W. (2019). Flanking heterozygosity influences the relative probability of different base substitutions in humans. *Royal Society open science*, 6(9):191018.

[10] Arnheim, N. and Calabrese, P. (2009). Understanding what determines the frequency and pattern of human germline mutations. *Nature Reviews Genetics*, 10(7):478–488.

[11] Beal, M. A., Yauk, C. L., and Marchetti, F. (2017). From sperm to offspring: Assessing the heritable genetic consequences of paternal smoking and potential public health impacts. *Mutation Research/Reviews in Mutation Research*, 773:26–50.

[12] Bergstrom, E. N., Huang, M. N., Mahto, U., Barnes, M., Stratton, M. R., Rozen, S. G., and Alexandrov, L. B. (2019). Sigprofilermatrixgenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics*, 20(1):1–12.

[13] Bertl, J., Guo, Q., Juul, M., Besenbacher, S., Nielsen, M. M., Hornshøj, H., Pedersen, J. S., and Hobolth, A. (2018). A site specific model and analysis of the neutral somatic mutation rate in whole-genome cancer data. *BMC Bioinformatics*, 19(1):1–15.

[14] Besenbacher, S., Sulem, P., Helgason, A., Helgason, H., Kristjansson, H., Jonasdottir, A., Jonasdottir, A., Magnusson, O. T., Thorsteinsdottir, U., Masson, G., et al. (2016). Multi-nucleotide de novo mutations in humans. *PLoS Genetics*, 12(11):e1006315.

[15] Bhérer, C., Campbell, C. L., and Auton, A. (2017). Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nature Communications*, 8(1):1–9.

[16] Blake, R., Hess, S. T., and Nicholson-Tuell, J. (1992). The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *Journal of Molecular Evolution*, 34(3):189–200.

[17] Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, 33(3):228–237.

[18] Boycott, K. M., Vanstone, M. R., Bulman, D. E., and MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics*, 14(10):681–691.

[19] BrainSpan (2016). BrainSpan: Atlas of the developing human brain.

[20] Campbell, I. M., Yuan, B., Robberecht, C., Pfundt, R., Szafranski, P., McEntagart, M. E., Nagamani, S. C., Erez, A., Bartnik, M., Wiśniowiecka-Kowalnik, B., et al. (2014). Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. *The American Journal of Human Genetics*, 95(2):173–182.

[21] Campos, D., Matos, S., and Oliveira, J. L. (2013). A modular framework for biomedical concept recognition. *BMC Bioinformatics*, 14(1):1–21.

[22] Carlson, J., Locke, A. E., Flickinger, M., Zawistowski, M., Levy, S., Myers, R. M., Boehnke, M., Kang, H. M., Scott, L. J., Li, J. Z., et al. (2018). Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nature Communications*, 9(1):1–13.

[23] Carter, T., Lyon, M. F., and Phillips, R. J. (1956). Induction of mutations in mice by chronic gamma irradiation; interim report. *The British Journal of Radiology*, 29(338):106–108.

[24] Cassa, C. A., Weghorn, D., Balick, D. J., Jordan, D. M., Nusinow, D., Samocha, K. E., O'Donnell-Luria, A., MacArthur, D. G., Daly, M. J., Beier, D. R., et al. (2017). Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nature Genetics*, 49(5):806–810.

[25] Cawthon, R. M., Meeks, H. D., Sasani, T. A., Smith, K. R., Kerber, R. A., O'Brien, E., Baird, L., Dixon, M. M., Peiffer, A. P., Leppert, M. F., et al. (2020). Germline mutation rates in young adults predict longevity and reproductive lifespan. *medRxiv*, page 19004184.

[26] Chance, P. F., Alderson, M. K., Leppig, K. A., Lensch, M. W., Matsunami, N., Smith, B., Swanson, P. D., Odelberg, S. J., Disteche, C. M., and Bird, T. D. (1993). DNA deletion associated with hereditary neuropathy with liability to pressure palsies. *Cell*, 72(1):143–151.

[27] Chen, J.-M., Férec, C., and Cooper, D. N. (2015). Complex multiple-nucleotide substitution mutations causing human inherited disease reveal novel insights into the action of translesion synthesis DNA polymerases. *Human Mutation*, 36(11):1034–1038.

[28] Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, 45(10):1127.

[29] Cleaver, J. E., Lam, E. T., and Revet, I. (2009). Disorders of nucleotide excision repair: the genetic and molecular basis of heterogeneity. *Nature Reviews Genetics*, 10(11):756–768.

[30] Clement, Y. and Arndt, P. F. (2011). Substitution patterns are under different influences in primates and rodents. *Genome Biology and Evolution*, 3:236–245.

[31] Coe, B. P., Stessman, H. A., Sulovari, A., Geisheker, M. R., Bakken, T. E., Lake, A. M., Dougherty, J. D., Lein, E. S., Hormozdiari, F., Bernier, R. A., et al. (2019). Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nature Genetics*, 51(1):106–116.

[32] Coe, B. P., Witherspoon, K., Rosenfeld, J. A., Van Bon, B. W., Vulto-van Silfhout, A. T., Bosco, P., Friend, K. L., Baker, C., Buono, S., Vissers, L. E., et al. (2014). Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nature Genetics*, 46(10):1063.

[33] Conrad, D. F., Keebler, J. E., DePristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C. L., Torroja, C., Garimella, K. V., et al. (2011). Variation in genome-wide mutation rates within and between human families. *Nature Genetics*, 43(7):712.

[34] Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

[35] Consortium, M. G. S., Waterston, R., Lindblad-Toh, K., Birney, E., and Rogers, J. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562.

[36] Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., et al. (2011). A copy number variation morbidity map of developmental delay. *Nature Genetics*, 43(9):838.

[37] Coulondre, C., Miller, J. H., Farabaugh, P. J., and Gilbert, W. (1978). Molecular basis of base substitution hotspots in escherichia coli. *Nature*, 274(5673):775–780.

[38] Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., et al. (2014). The reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1):D472–D477.

[39] De Ligt, J., Willemsen, M. H., Van Bon, B. W., Kleefstra, T., Yntema, H. G., Kroes, T., Vulto-van Silfhout, A. T., Koolen, D. A., De Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. *New England Journal of Medicine*, 367(20):1921–1929.

[40] de Vries, B. B., Pfundt, R., Leisink, M., Koolen, D. A., Vissers, L. E., Janssen, I. M., van Reijmersdal, S., Nillesen, W. M., Huys, E. H., de Leeuw, N., et al. (2005). Diagnostic genome profiling in mental retardation. *The American Journal of Human Genetics*, 77(4):606–616.

[41] Deciphering Developmental Disorders Study (2017). Prevalence and architecture of de novo mutations in developmental disorders. *Nature*, 542(7642):433.

[42] Deelen, P., van Dam, S., Herkert, J. C., Karjalainen, J. M., Brugge, H., Abbott, K. M., van Diemen, C. C., van der Zwaag, P. A., Gerkes, E. H., Zonneveld-Huijssoon, E., et al. (2019). Improving the diagnostic yield of exome-sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. *Nature Communications*, 10(1):1–13.

[43] DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491.

[44] Diaz, J., Berger, S., and Leon, E. (2019). TFE3-associated neurodevelopmental disorder: A distinct recognizable syndrome. *American Journal of Medical Genetics Part A*.

[45] Duncan, B. K. and Miller, J. H. (1980). Mutagenic deamination of cytosine residues in DNA. *Nature*, 287(5782):560–561.

[46] Dutil, J., Teer, J. K., Golubeva, V., Yoder, S., Tong, W. L., Arroyo, N., Karam, R., Echenique, M., Matta, J. L., and Monteiro, A. N. (2019). Germline variants in cancer genes in high-risk non-brca patients from puerto rico. *Scientific Reports*, 9(1):1–11.

[47] Elder, R. H., Jansen, J. G., Weeks, R. J., Willington, M. A., Deans, B., Watson, A. J., Mynett, K. J., Bailey, J. A., Cooper, D. P., Rafferty, J. A., et al. (1998). Alkylpurine–dna–n-glycosylase knockout mice show increased susceptibility to induction of mutations by methyl methanesulfonate. *Molecular and Cellular Biology*, 18(10):5828–5837.

[48] Eyler, D. E., Burnham, K. A., Wilson, T. E., and O’Brien, P. J. (2017). Mechanisms of glycosylase induced genomic instability. *PloS One*, 12(3).

[49] Fan, H. and Chu, J.-Y. (2007). A brief review of short tandem repeat mutation. *Genomics, Proteomics & Bioinformatics*, 5(1):7–14.

[50] Firth, H. V. and Wright, C. F. (2011). The Deciphering Developmental Disorders (DDD) study. *Developmental Medicine & Child Neurology*, 53(8):702–703.

[51] Forster, L., Forster, P., Lutz-Bonengel, S., Willkomm, H., and Brinkmann, B. (2002). Natural radioactivity and human mitochondrial DNA mutations. *Proceedings of the National Academy of Sciences*, 99(21):13950–13954.

[52] Francioli, L. C., Polak, P. P., Koren, A., Menelaou, A., Chun, S., Renkens, I., Van Duijn, C. M., Swertz, M., Wijmenga, C., Van Ommen, G., et al. (2015). Genome-wide patterns and properties of de novo mutations in humans. *Nature Genetics*, 47(7):822.

[53] Fu, W., Zhang, F., Wang, Y., Gu, X., and Jin, L. (2010). Identification of copy number variation hotspots in human populations. *The American Journal of Human Genetics*, 87(4):494–504.

[54] Gao, Z., Moorjani, P., Sasani, T. A., Pedersen, B. S., Quinlan, A. R., Jorde, L. B., Amster, G., and Przeworski, M. (2019). Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proceedings of the National Academy of Sciences*, 116(19):9491–9500.

[55] Garcia-Diaz, M. and Kunkel, T. A. (2006). Mechanism of a genetic glissando*: structural biology of indel mutations. *Trends in Biochemical Sciences*, 31(4):206–214.

[56] Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

[57] Gerstung, M., Papaemmanuil, E., and Campbell, P. J. (2014). Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics*, 30(9):1198–1204.

[58] Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., van de Vorst, M., van Bon, B. W., Willemsen, M. H., Kwint, M., Janssen, I. M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511(7509):344–347.

[59] Gilissen, C., Hoischen, A., Brunner, H. G., and Veltman, J. A. (2012). Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics*, 20(5):490–497.

[60] Girard, S. L., Bourassa, C. V., Lemieux Perreault, L.-P., Legault, M.-A., Barhdadi, A., Ambalavanan, A., Brendgen, M., Vitaro, F., Noreau, A., Dionne, G., et al. (2016). Paternal age explains a major portion of de novo germline mutation rate variability in healthy individuals. *PLoS One*, 11(10):e0164212.

[61] Glassner, B. J., Rasmussen, L. J., Najarian, M. T., Posnick, L. M., and Samson, L. D. (1998). Generation of a strong mutator phenotype in yeast by imbalanced base excision repair. *Proceedings of the National Academy of Sciences*, 95(17):9997–10002.

[62] Goldmann, J. M., Wong, W. S., Pinelli, M., Farrah, T., Bodian, D., Stittrich, A. B., Glusman, G., Vissers, L. E., Hoischen, A., Roach, J. C., et al. (2016). Parent-of-origin-specific signatures of de novo mutations. *Nature Genetics*, 48(8):935.

[63] Goodman, M. (1963). Serological analysis of the systematics of recent hominoids. *Human Biology*, 35(3):377–436.

[64] Goriely, A., Hansen, R. M., Taylor, I. B., Olesen, I. A., Jacobsen, G. K., McGowan, S. J., Pfeifer, S. P., McVean, G. A., Rajpert-De Meyts, E., and Wilkie, A. O. (2009). Activating mutations in FGFR3 and hras reveal a shared genetic origin for congenital disorders and testicular tumors. *Nature Genetics*, 41(11):1247.

[65] Goriely, A., McGrath, J. J., Hultman, C. M., Wilkie, A. O., and Malaspina, D. (2013). "selfish spermatogonial selection": a novel mechanism for the association between advanced paternal age and neurodevelopmental disorders. *American Journal of Psychiatry*, 170(6):599–608.

[66] Goriely, A. and Wilkie, A. O. (2010). Missing heritability: paternal age effect mutations and selfish spermatogonia. *Nature Reviews Genetics*, 11(8):589–589.

[67] Goriely, A. and Wilkie, A. O. (2012). Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *The American Journal of Human Genetics*, 90(2):175–200.

[68] Green, P., Ewing, B., Miller, W., Thomas, P. J., and Green, E. D. (2003). Transcription-associated mutational asymmetry in mammalian evolution. *Nature Genetics*, 33(4):514–517.

[69] Guo, H., Wang, T., Wu, H., Long, M., Coe, B. P., Li, H., Xun, G., Ou, J., Chen, B., Duan, G., et al. (2018). Inherited and multiple de novo mutations in autism/developmental delay risk genes suggest a multifactorial model. *Molecular Autism*, 9(1):64.

[70] Gymrek, M., Willems, T., Reich, D., and Erlich, Y. (2017). Interpreting short tandem repeat variations in humans using mutational constraint. *Nature Genetics*, 49(10):1495.

[71] Haldane, J. (1946). The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Annals of Eugenics*, 13(1):262–271.

[72] Haldane, J. B. (1935). The rate of spontaneous mutation of a human gene. *Journal of Genetics*, 31(3):317.

[73] Halldorsson, B. V., Palsson, G., Stefansson, O. A., Jonsson, H., Hardarson, M. T., Eggertsson, H. P., Gunnarsson, B., Oddsson, A., Halldorsson, G. H., Zink, F., et al. (2019). Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science*, 363(6425):eaau1043.

[74] Happle, R. (1987). Lethal genes surviving by mosaicism: a possible explanation for sporadic birth defects involving the skin. *Journal of the American Academy of Dermatology*, 16(4):899–906.

[75] Harris, K. (2015). Evidence for recent, population-specific evolution of the human mutation rate. *Proceedings of the National Academy of Sciences*, 112(11):3439–3444.

[76] Harris, K. and Nielsen, R. (2014). Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Research*, 24(9):1445–1454.

[77] Harris, R. S. (2013). Cancer mutation signatures, DNA damage mechanisms, and potential clinical implications. *Genome Medicine*, 5(9):87.

[78] He, X., Sanders, S. J., Liu, L., De Rubeis, S., Lim, E. T., Sutcliffe, J. S., Schellenberg, G. D., Gibbs, R. A., Daly, M. J., Buxbaum, J. D., et al. (2013). Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genetics*, 9(8).

[79] Hendrich, B., Hardeland, U., Ng, H.-H., Jiricny, J., and Bird, A. (1999). The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature*, 401(6750):301–304.

[80] Hodgkinson, A. and Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*, 12(11):756–766.

[81] Hu, C., Hart, S. N., Polley, E. C., Gnanaolivu, R., Shimelis, H., Lee, K. Y., Lilyquist, J., Na, J., Moore, R., Antwi, S. O., et al. (2018). Association between inherited germline mutations in cancer predisposition genes and risk of pancreatic cancer. *JAMA*, 319(23):2401–2409.

[82] Huang, K.-l., Mashl, R. J., Wu, Y., Ritter, D. I., Wang, J., Oh, C., Paczkowska, M., Reynolds, S., Wyczalkowski, M. A., Oak, N., et al. (2018). Pathogenic germline variants in 10,389 adult cancers. *Cell*, 173(2):355–370.

[83] Huang, N., Lee, I., Marcotte, E. M., and Hurles, M. E. (2010). Characterising and predicting haploinsufficiency in the human genome. *PLoS Genetics*, 6(10).

[84] Huang, Q.-Y., Xu, F.-H., Shen, H., Deng, H.-Y., Liu, Y.-J., Liu, Y.-Z., Li, J.-L., Recker, R. R., and Deng, H.-W. (2002). Mutation patterns at dinucleotide microsatellite loci in humans. *The American Journal of Human Genetics*, 70(3):625–634.

[85] Huang, Y.-F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature Genetics*, 49(4):618–624.

[86] Hwang, D. G. and Green, P. (2004). Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences*, 101(39):13994–14001.

[87] Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics*, 36(9):949–951.

[88] Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics*, 48(2):214.

[89] Iossifov, I., O'roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, 515(7526):216–221.

[90] Itsara, A., Wu, H., Smith, J. D., Nickerson, D. A., Romieu, I., London, S. J., and Eichler, E. E. (2010). De novo rates and selection of large copy number variation. *Genome Research*, 20(11):1469–1481.

[91] Jaganathan, K., Panagiotopoulou, S. K., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., et al. (2019). Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548.

[92] Jager, M., Blokzijl, F., Kuijk, E., Bertl, J., Vougioukalaki, M., Janssen, R., Besselink, N., Boymans, S., de Ligt, J., Pedersen, J. S., et al. (2019). Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. *Genome Research*, 29(7):1067–1077.

[93] Jeffreys, A. J., Royle, N. J., Wilson, V., and Wong, Z. (1988). Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature*, 332(6161):278–281.

[94] Jeffreys, A. J., Tamaki, K., MacLeod, A., Monckton, D. G., Neil, D. L., and Armour, J. A. (1994). Complex gene conversion events in germline mutation at human minisatellites. *Nature Genetics*, 6(2):136–145.

[95] Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985). Hypervariable 'minisatellite'regions in human DNA. *Nature*, 314(6006):67–73.

[96] Johnson, R. E., Washington, M. T., Prakash, S., and Prakash, L. (2000). Fidelity of human DNA polymerase η. *Journal of Biological Chemistry*, 275(11):7447–7450.

[97] Jónsson, H., Sulem, P., Arnadottir, G. A., Pálsson, G., Eggertsson, H. P., Kristmundsdottir, S., Zink, F., Kehr, B., Hjorleifsson, K. E., Jensson, B. Ö., et al. (2018). Multiple transmissions of de novo mutations in families. *Nature Genetics*, 50(12):1674–1680.

[98] Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M. T., Hjorleifsson, K. E., Eggertsson, H. P., Gudjonsson, S. A., et al. (2017). Parental influence on human germline de novo mutations in 1,548 trios from iceland. *Nature*, 549(7673):519–522.

[99] Ju, Y. S., Martincorena, I., Gerstung, M., Petljak, M., Alexandrov, L. B., Rahbari, R., Wedge, D. C., Davies, H. R., Ramakrishna, M., Fullam, A., et al. (2017). Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature*, 543(7647):714–718.

[100] Kaplanis, J., Akawi, N., Gallone, G., McRae, J. F., Prigmore, E., Wright, C. F., Fitzpatrick, D. R., Firth, H. V., Barrett, J. C., Hurles, M. E., et al. (2019). Exome-wide assessment of the functional impact and pathogenicity of multinucleotide mutations. *Genome Research*, 29(7):1047–1056.

[101] Kaplanis, J., Samocha, K. E., Wiel, L., Zhang, Z., Arvai, K. J., Eberhardt, R. Y., Gallone, G., Lelieveld, S. H., Martin, H. C., McRae, J. F., et al. (2020). Integrating healthcare and research genetic data empowers the discovery of 28 novel developmental disorders. *BioRxiv*, page 797787.

[102] Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, page 531210.

[103] Kato, T., Inagaki, H., Yamada, K., Kogo, H., Ohye, T., Kowa, H., Nagaoka, K., Taniguchi, M., Emanuel, B. S., and Kurahashi, H. (2006). Genetic variation affects de novo translocation frequency. *Science*, 311(5763):971–971.

[104] Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human protein reference database—2009 update. *Nucleic Acids Research*, 37(suppl_1):D767–D772.

[105] Kilpivaara, O. and Aaltonen, L. (2013). Diagnostic cancer genome sequencing and the contribution of germline variants. *Science*, 339(6127):1559–1562.

[106] Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626.

[107] Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310.

[108] Klein, H. L. (2017). Genome instabilities arising from ribonucleotides in DNA. *DNA Repair*, 56:26–32.

[109] Kondrashov, A. S. (2003). Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Human Mutation*, 21(1):12–27.

[110] Kondrashov, A. S. and Crow, J. F. (1993). A molecular approach to estimating the human deleterious mutation rate. *Human mutation*, 2(3):229–234.

[111] Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412):471–475.

[112] Kosmicki, J. A., Samocha, K. E., Howrigan, D. P., Sanders, S. J., Slowikowski, K., Lek, M., Karczewski, K. J., Cutler, D. J., Devlin, B., Roeder, K., et al. (2017). Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nature Genetics*, 49(4):504.

[113] Kurek, K. C., Luks, V. L., Ayturk, U. M., Alomari, A. I., Fishman, S. J., Spencer, S. A., Mulliken, J. B., Bowen, M. E., Yamamoto, G. L., Kozakewich, H. P., et al. (2012). Somatic mosaic activating mutations in PIK3CA cause CLOVES syndrome. *The American Journal of Human Genetics*, 90(6):1108–1115.

[114] Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J. C., and Dry, J. R. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Research*, 44(11):e108–e108.

[115] Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., and Maglott, D. R. (2013). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1):D980–D985.

[116] Lange, S. S., Takata, K.-i., and Wood, R. D. (2011). DNA polymerases and cancer. *Nature Reviews Cancer*, 11(2):96–110.

[117] Laskowski, R. A., Stephenson, J. D., Sillitoe, I., Orengo, C. A., and Thornton, J. M. (2020). VarSite: Disease variants and protein structure. *Protein Science*, 29(1):111–119.

[118] Lau, A. Y., Wyatt, M. D., Glassner, B. J., Samson, L. D., and Ellenberger, T. (2000). Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, aag. *Proceedings of the National Academy of Sciences*, 97(25):13573–13578.

[119] Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501.

[120] Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.

[121] Leclercq, S., Rivals, E., and Jarne, P. (2010). DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. *Genome Biology and Evolution*, 2:325–335.

[122] Leeksma, O., De Miranda, N., and Veelken, H. (2017). Germline mutations predisposing to diffuse large b-cell lymphoma. *Blood Cancer Journal*, 7(2):e532–e532.

[123] Lehmann, A. R., McGibbon, D., and Stefanini, M. (2011). Xeroderma pigmentosum. *Orphanet Journal of Rare Diseases*, 6(1):70.

[124] Lejeune, J., Turpin, R., and Gautier, M. (1959). Le mongolisme, premier exemple d'aberration autosomique humaine. *Ann Genet*, 1(4):1–49.

[125] Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285.

[126] Lenz, C., Haerty, W., and Golding, G. B. (2014). Increased substitution rates surrounding low-complexity regions within primate proteins. *Genome Biology and Evolution*, 6(3):655–665.

[127] Lercher, M. J. and Hurst, L. D. (2002). Human SNP variability and mutation rate are higher in regions of high recombination. *Trends in Genetics*, 18(7):337–340.

[128] Li, C. and Luscombe, N. M. (2019). Nucleosome positioning stability is a significant modulator of germline mutation rate variation across the human genome. *BioRxiv*.

[129] Li, W.-H., Wu, C.-I., and Luo, C.-C. (1984). Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *Journal of Molecular Evolution*, 21(1):58–71.

[130] Lindsay, S. J., Rahbari, R., Kaplanis, J., Keane, T., and Hurles, M. E. (2019). Similarities and differences in patterns of germline mutation between mice and humans. *Nature Communications*, 10(1):1–12.

[131] Liu, P., Yuan, B., Carvalho, C. M., Wuster, A., Walter, K., Zhang, L., Gambin, T., Chong, Z., Campbell, I. M., Akdemir, Z. C., et al. (2017). An organismal CNV mutator phenotype restricted to early human development. *Cell*, 168(5):830–842.

[132] Longley, M. J., Nguyen, D., Kunkel, T. A., and Copeland, W. C. (2001). The fidelity of human DNA polymerase γ with and without exonucleolytic proofreading and the p55 accessory subunit. *Journal of Biological Chemistry*, 276(42):38555–38562.

[133] Lord, J., McMullan, D. J., Eberhardt, R. Y., Rinck, G., Hamilton, S. J., Quinlan-Jones, E., Prigmore, E., Keelagher, R., Best, S. K., Carey, G. K., et al. (2019). Prenatal exome sequencing analysis in fetal structural anomalies detected by ultrasonography (PAGE): a cohort study. *The Lancet*, 393(10173):747–757.

[134] Lupski, J. R., de Oca-Luna, R. M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B. J., Saucedo-Cardenas, O., Barker, D. F., Killian, J. M., Garcia, C. A., et al. (1991). DNA duplication associated with charcot-marie-tooth disease type 1a. *Cell*, 66(2):219–232.

[135] Lynch, M. (2010). Evolution of the mutation rate. *TRENDS in Genetics*, 26(8):345–352.

[136] Maher, G. J., McGowan, S. J., Giannoulatou, E., Verrill, C., Goriely, A., and Wilkie, A. O. (2016). Visualizing the origins of selfish de novo mutations in individual seminiferous tubules of human testes. *Proceedings of the National Academy of Sciences*, 113(9):2454–2459.

[137] Maher, G. J., Ralph, H. K., Ding, Z., Koelling, N., Mlcochova, H., Giannoulatou, E., Dhami, P., Paul, D. S., Stricker, S. H., Beck, S., et al. (2018). Selfish mutations dysregulating RAS-MAPK signaling are pervasive in aged human testes. *Genome Research*, 28(12):1779–1790.

[138] Makova, K. D. and Hardison, R. C. (2015). The effects of chromatin organization on variation in mutation rates in the genome. *Nature Reviews Genetics*, 16(4):213–223.

[139] Marchetti, F., Rowan-Carroll, A., Williams, A., Polyzos, A., Berndt-Weis, M. L., and Yauk, C. L. (2011). Sidestream tobacco smoke is a male germ cell mutagen. *Proceedings of the National Academy of Sciences*, 108(31):12811–12814.

[140] Martin, H. C., Jones, W. D., McIntyre, R., Sanchez-Andrade, G., Sanderson, M., Stephenson, J. D., Jones, C. P., Handsaker, J., Gallone, G., Bruntraeger, M., et al. (2018). Quantifying the contribution of recessive coding variation to developmental disorders. *Science*, 362(6419):1161–1164.

[141] Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., and Campbell, P. J. (2017). Universal patterns of selection in cancer and somatic tissues. *Cell*, 171(5):1029–1041.

[142] Mathieson, I. and Reich, D. (2017). Differences in the rare variant spectrum among human populations. *PLoS Genetics*, 13(2):e1006581.

[143] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

[144] McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. *Genome Biology*, 17(1):122.

[145] Miccoli, L., Burr, K. L., Hickenbotham, P., Friedberg, E. C., Angulo, J. F., and Dubrova, Y. E. (2007). The combined effects of xeroderma pigmentosum c deficiency and mutagens on mutation rates in the mouse germ line. *Cancer Research*, 67(10):4695–4699.

[146] Michaelson, J. J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A., et al. (2012). Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*, 151(7):1431–1442.

[147] Millar, C. B., Guy, J., Sansom, O. J., Selfridge, J., MacDougall, E., Hendrich, B., Keightley, P. D., Bishop, S. M., Clarke, A. R., and Bird, A. (2002). Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. *Science*, 297(5580):403–405.

[148] Miller, J. A., Ding, S.-L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., Ebbert, A., Riley, Z. L., Royall, J. J., Aiona, K., et al. (2014). Transcriptional landscape of the prenatal human brain. *Nature*, 508(7495):199–206.

[149] Mills, R. E., Pittard, W. S., Mullaney, J. M., Farooq, U., Creasy, T. H., Mahurkar, A. A., Kemeza, D. M., Strassler, D. S., Ponting, C. P., Webber, C., et al. (2011). Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Research*, 21(6):830–839.

[150] Monckton, D. G., Neumann, R., Guram, T., Fretwell, N., Tamaki, K., MacLeod, A., and Jeffreys, A. J. (1994). Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. *Nature Genetics*, 8(2):162–170.

[151] Montgomery, S. B., Goode, D. L., Kvikstad, E., Albers, C. A., Zhang, Z. D., Mu, X. J., Ananda, G., Howie, B., Karczewski, K. J., Smith, K. S., et al. (2013). The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Research*, 23(5):749–761.

[152] Moorjani, P., Amorim, C. E. G., Arndt, P. F., and Przeworski, M. (2016). Variation in the molecular clock of primates. *Proceedings of the National Academy of Sciences*, 113(38):10607–10612.

[153] Mugal, C. F., von Grünberg, H.-H., and Peifer, M. (2009). Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Molecular Biology and Evolution*, 26(1):131–142.

[154] Muller, H. J. (1928). The measurement of gene mutation rate in drosophila, its high variability, and its dependence upon temperature. *Genetics*, 13(4):279.

[155] Nachman, M. W. (2004). Haldane and the first estimates of the human mutation rate. *Journal of Genetics*, 83(3):231–233.

[156] Nachman, M. W. and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304.

[157] Neale, B. M., Kou, Y., Liu, L., Ma'Ayan, A., Samocha, K. E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397):242–245.

[158] Neveling, K., Feenstra, I., Gilissen, C., Hoefsloot, L. H., Kamsteeg, E.-J., Mensenkamp, A. R., Rodenburg, R. J., Yntema, H. G., Spruijt, L., Vermeer, S., et al. (2013). A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. *Human Mutation*, 34(12):1721–1726.

[159] Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993.

[160] Ohashi, E., Ogi, T., Kusumoto, R., Iwai, S., Masutani, C., Hanaoka, F., and Ohmori, H. (2000). Error-prone bypass of certain DNA lesions by the human DNA polymerase κ. *Genes & Development*, 14(13):1589–1594.

[161] Okajima, K., Warman, M. L., Byrne, L. C., and Kerr, D. S. (2006). Somatic mosaicism in a male with an exon skipping mutation in PDHA1 of the pyruvate dehydrogenase complex results in a milder phenotype. *Molecular Genetics and Metabolism*, 87(2):162–168.

[162] O'roak, B., Stessman, H., Boyle, E., Witherspoon, K., Martin, B., Lee, C., Vives, L., Baker, C., Hiatt, J., Nickerson, D., et al. (2014). Recurrent de novo mutations implicate novel genes underlying simplex autism risk. *Nature Communications*, 5(1):1–6.

[163] O'Roak, B. J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J. J., Girirajan, S., Karakoc, E., MacKenzie, A. P., Ng, S. B., Baker, C., et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics*, 43(6):585.

[164] Petljak, M., Alexandrov, L. B., Brammeld, J. S., Price, S., Wedge, D. C., Grossmann, S., Dawson, K. J., Ju, Y. S., Iorio, F., Tubio, J. M., et al. (2019). Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell*, 176(6):1282–1294.

[165] Phillips, D. H. (2018). Mutational spectra and mutational signatures: Insights into cancer aetiology and mechanisms of DNA damage and repair. *DNA Repair*, 71:6–11.

[166] Pich, O., Muiños, F., Lolkema, M. P., Steeghs, N., Gonzalez-Perez, A., and Lopez-Bigas, N. (2019). The mutational footprints of cancer therapies. *Nature Genetics*, 51(12):1732–1740.

[167] Pino, M. S., Mino-Kenudson, M., Wildemore, B. M., Ganguly, A., Batten, J., Sperduti, I., Iafrate, A. J., and Chung, D. C. (2009). Deficient DNA mismatch repair is common in Lynch syndrome-associated colorectal adenomas. *The Journal of Molecular Diagnostics*, 11(3):238–247.

[168] Pinto, Y., Gabay, O., Arbiza, L., Sams, A. J., Keinan, A., and Levanon, E. Y. (2016). Clustered mutations in hominid genome evolution are consistent with APOBEC3G enzymatic activity. *Genome Research*, 26(5):579–587.

[169] Pippard, E. C., Hall, A. J., Barker, D. J., and Bridges, B. A. (1988). Cancer in homozygotes and heterozygotes of ataxia-telangiectasia and xeroderma pigmentosum in britain. *Cancer Research*, 48(10):2929–2932.

[170] Plant, K. E., Boye, E., Green, P. M., Vetrie, D., and Flinter, F. A. (2000). Somatic mosaicism associated with a mild Alport syndrome phenotype. *Journal of Medical Genetics*, 37(3):238–239.

[171] Polak, P., Karlić, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M. S., Reynolds, A., Rynes, E., Vlahoviček, K., Stamatoyannopoulos, J. A., et al. (2015). Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, 518(7539):360–364.

[172] Porta-Pardo, E., Kamburov, A., Tamborero, D., Pons, T., Grases, D., Valencia, A., Lopez-Bigas, N., Getz, G., and Godzik, A. (2017). Comparison of algorithms for the detection of cancer drivers at subgene resolution. *Nature Methods*, 14(8):782–788.

[173] Rahbari, R., Wuster, A., Lindsay, S. J., Hardwick, R. J., Alexandrov, L. B., Al Turki, S., Dominiczak, A., Morris, A., Porteous, D., Smith, B., et al. (2016). Timing, rates and spectra of human germline mutation. *Nature Genetics*, 48(2):126.

[174] Ramu, A., Noordam, M. J., Schwartz, R. S., Wuster, A., Hurles, M. E., Cartwright, R. A., and Conrad, D. F. (2013). DeNovoGear: de novo indel and point mutation discovery and phasing. *Nature Methods*, 10(10):985.

[175] Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., et al. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444–454.

[176] Reijns, M. A., Rabe, B., Rigby, R. E., Mill, P., Astell, K. R., Lettice, L. A., Boyle, S., Leitch, A., Keighren, M., Kilanowski, F., et al. (2012). Enzymatic removal of ribonucleotides from DNA is essential for mammalian genome integrity and development. *Cell*, 149(5):1008–1022.

[177] Retterer, K., Juusola, J., Cho, M. T., Vitazka, P., Millan, F., Gibellini, F., Vertino-Bell, A., Smaoui, N., Neidich, J., Monaghan, K. G., et al. (2016). Clinical application of whole-exome sequencing across clinical indications. *Genetics in Medicine*, 18(7):696–704.

[178] Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R., Wilkie, A. O., McVean, G., Lunter, G., Consortium, W., et al. (2014). Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, 46(8):912.

[179] Rizzato, F., Rodriguez, A., and Laio, A. (2016). Non-markovian effects on protein sequence evolution due to site dependent substitution rates. *BMC Bioinformatics*, 17(1):258.

[180] Roach, J. C., Glusman, G., Smit, A. F., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., et al. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978):636–639.

[181] Roberts, S. A., Lawrence, M. S., Klimczak, L. J., Grimm, S. A., Fargo, D., Stojanov, P., Kiezun, A., Kryukov, G. V., Carter, S. L., Saksena, G., et al. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature Genetics*, 45(9):970.

[182] Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26.

[183] Ropers, H. H. (2010). Genetics of early onset cognitive impairment. *Annual Review of Genomics and Human Genetics*, 11:161–187.

[184] Russell, W. (1951). X-ray-induced mutations in mice. In *Cold Spring Harbor symposia on quantitative biology*, volume 16, pages 327–336. Cold Spring Harbor Laboratory Press.

[185] Russell, W. (1989). Reminiscences of a mouse specific-locus test addict. *Environmental and Molecular Mutagenesis*, 14(S16):16–22.

[186] Samocha, K. E., Kosmicki, J. A., Karczewski, K. J., O'Donnell-Luria, A. H., Pierce-Hoffman, E., MacArthur, D. G., Neale, B. M., and Daly, M. J. (2017). Regional missense constraint improves variant deleteriousness prediction. *BioRxiv*, page 148353.

[187] Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, 46(9):944.

[188] Sandmann, S., De Graaf, A. O., Karimi, M., Van Der Reijden, B. A., Hellström-Lindberg, E., Jansen, J. H., and Dugas, M. (2017). Evaluating variant calling tools for non-matched next-generation sequencing data. *Scientific Reports*, 7:43169.

[189] Sasani, T. A., Pedersen, B. S., Gao, Z., Baird, L., Przeworski, M., Jorde, L. B., and Quinlan, A. R. (2019). Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *elife*, 8.

[190] Scally, A. and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics*, 13(10):745–753.

[191] Schrider, D. R., Hourmozdi, J. N., and Hahn, M. W. (2011). Pervasive multinucleotide mutational events in eukaryotes. *Current Biology*, 21(12):1051–1054.

[192] Searle, A. (1984). The specific locus test in the mouse. In *Handbook of mutagenicity test procedures*, pages 371–391. Elsevier Amsterdam.

[193] Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong association of de novo copy number mutations with autism. *Science*, 316(5823):445–449.

[194] Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., et al. (2004). Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528.

[195] Ségurel, L., Wyman, M. J., and Przeworski, M. (2014). Determinants of mutation rate variation in the human germline. *Annual Review of Genomics and Human Genetics*, 15:47–70.

[196] Seidman, M. M., Bredberg, A., Seetharam, S., and Kraemer, K. H. (1987). Multiple point mutations in a shuttle vector propagated in human cells: evidence for an error-prone DNA polymerase activity. *Proceedings of the National Academy of Sciences*, 84(14):4944–4948.

[197] Seoighe, C. and Scally, A. (2017). Inference of candidate germline mutator loci in humans from genome-wide haplotype data. *PLoS Genetics*, 13(1):e1006549.

[198] Sheridan, E., Wright, J., Small, N., Corry, P. C., Oddie, S., Whibley, C., Petherick, E. S., Malik, T., Pawson, N., McKinney, P. A., et al. (2013). Risk factors for congenital anomaly in a multiethnic birth cohort: an analysis of the born in bradford study. *The Lancet*, 382(9901):1350–1359.

[199] Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N., Gaunt, T. R., and Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, 31(10):1536–1543.

[200] Shlien, A., Tabori, U., Marshall, C. R., Pienkowska, M., Feuk, L., Novokmet, A., Nanda, S., Druker, H., Scherer, S. W., and Malkin, D. (2008). Excessive genomic DNA copy number variation in the Li–Fraumeni cancer predisposition syndrome. *Proceedings of the National Academy of Sciences*, 105(32):11264–11269.

[201] Short, P. J., McRae, J. F., Gallone, G., Sifrim, A., Won, H., Geschwind, D. H., Wright, C. F., Firth, H. V., FitzPatrick, D. R., Barrett, J. C., et al. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature*, 555(7698):611–616.

[202] Singh, T., Walters, J. T., Johnstone, M., Curtis, D., Suvisaari, J., Torniainen, M., Rees, E., Iyegbe, C., Blackwood, D., McIntosh, A. M., et al. (2017). The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. *Nature Genetics*, 49(8):1167.

[203] Smedley, D., Schubach, M., Jacobsen, J. O., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N. L., McMurry, J. A., et al. (2016). A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *The American Journal of Human Genetics*, 99(3):595–606.

[204] Smith, T. C., Arndt, P. F., and Eyre-Walker, A. (2018). Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. *PLoS Genetics*, 14(3):e1007254.

[205] Soemedi, R., Wilson, I. J., Bentham, J., Darlay, R., Töpf, A., Zelenika, D., Cosgrove, C., Setchfield, K., Thornborough, C., Granados-Riveron, J., et al. (2012). Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. *The American Journal of Human Genetics*, 91(3):489–501.

[206] Soojin, V. Y. (2013). Morris goodman's hominoid rate slowdown: The importance of being neutral. *Molecular Phylogenetics and Evolution*, 66(2):569–574.

[207] Stadler, L. (1932). On the genetic nature of induced mutations in plants, reprinted from the proceedings of the sixth international congress of genetics, vol. 1.

[208] Stamatoyannopoulos, J. A., Adzhubei, I., Thurman, R. E., Kryukov, G. V., Mirkin, S. M., and Sunyaev, S. R. (2009). Human mutation rate associated with DNA replication timing. *Nature Genetics*, 41(4):393.

[209] Stone, J. E., Lujan, S. A., and Kunkel, T. A. (2012). DNA polymerase zeta generates clustered mutations during bypass of endogenous DNA lesions in saccharomyces cerevisiae. *Environmental and Molecular Mutagenesis*, 53(9):777–786.

[210] Sun, J. X., Helgason, A., Masson, G., Ebenesersdóttir, S. S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., et al. (2012). A direct characterization of human mutation based on microsatellites. *Nature Genetics*, 44(10):1161.

[211] Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., et al. (2019). String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613.

[212] Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 47(D1):D941–D947.

[213] Tawn, E. J. (2002). Hereditary effects of radiation: Unscear 2001 report to the general assembly, with scientific annex.

[214] The, I., of Whole, T. P.-C. A., Consortium, G., et al. (2020). Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82.

[215] The Deciphering Developmental Disorder Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, 519(7542):223–228.

[216] Tissier, A., McDonald, J. P., Frank, E. G., and Woodgate, R. (2000). polι, a remarkably error-prone human DNA polymerase. *Genes & Development*, 14(13):1642–1650.

[217] Tubbs, A. and Nussenzweig, A. (2017). Endogenous DNA damage as a source of genomic instability in cancer. *Cell*, 168(4):644–656.

[218] Turner, D. J., Miretti, M., Rajan, D., Fiegler, H., Carter, N. P., Blayney, M. L., Beck, S., and Hurles, M. E. (2008). Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nature Genetics*, 40(1):90.

[219] Uchimura, A., Higuchi, M., Minakuchi, Y., Ohno, M., Toyoda, A., Fujiyama, A., Miura, I., Wakana, S., Nishino, J., and Yagi, T. (2015). Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome Research*, 25(8):1125–1134.

[220] Van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications*, 30(1):121–141.

[221] Veltman, J. A. and Brunner, H. G. (2012). De novo mutations in human genetic disease. *Nature Reviews Genetics*, 13(8):565–575.

[222] Venkat, A., Hahn, M. W., and Thornton, J. W. (2018). Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nature Ecology & Evolution*, 2(8):1280.

[223] Venkatarajan, M. S. and Braun, W. (2001). New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties. *Molecular Modeling Annual*, 7(12):445–453.

[224] Villegas, F., Lehalle, D., Mayer, D., Rittirsch, M., Stadler, M. B., Zinner, M., Olivieri, D., Vabres, P., Duplomb-Jego, L., De Bont, E. S., et al. (2019). Lysosomal signaling licenses embryonic stem cell differentiation via inactivation of TFE3. *Cell stem cell*, 24(2):257–270.

[225] Vissers, L. E., van Ravenswaaij, C. M., Admiraal, R., Hurst, J. A., de Vries, B. B., Janssen, I. M., van der Vliet, W. A., Huys, E. H., de Jong, P. J., Hamel, B. C., et al. (2004). Mutations in a new member of the chromodomain gene family cause charge syndrome. *Nature Genetics*, 36(9):955–957.

[226] Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276.

[227] Webb, A. J., Berg, I. L., and Jeffreys, A. (2008). Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. *Proceedings of the National Academy of Sciences*, 105(30):10471–10476.

[228] Wei, L., Liu, L. T., Conroy, J. R., Hu, Q., Conroy, J. M., Morrison, C. D., Johnson, C. S., Wang, J., and Liu, S. (2015). MAC: identifying and correcting annotation for multi-nucleotide variations. *BMC Genomics*, 16(1):569.

[229] Weinstein, L. S., Shenker, A., Gejman, P. V., Merino, M. J., Friedman, E., and Spiegel, A. M. (1991). Activating mutations of the stimulatory G protein in the McCune–Albright syndrome. *New England Journal of Medicine*, 325(24):1688–1695.

[230] Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., Telenti, A., and di Iulio, J. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nature Communications*, 10(1):1–9.

[231] Weren, R. D., Ligtenberg, M. J., Kets, C. M., De Voer, R. M., Verwiel, E. T., Spruijt, L., van Zelst-Stams, W. A., Jongmans, M. C., Gilissen, C., Hehir-Kwa, J. Y., et al. (2015). A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. *Nature Genetics*, 47(6):668.

[232] Wong, E., Yang, K., Kuraguchi, M., Werling, U., Avdievich, E., Fan, K., Fazzari, M., Jin, B., Brown, A. M., Lipkin, M., et al. (2002). MBD4 inactivation increases c>t transition mutations and promotes gastrointestinal tumor formation. *Proceedings of the National Academy of Sciences*, 99(23):14937–14942.

[233] Wong, J. T.-F. (1975). A co-evolution theory of the genetic code. *Proceedings of the National Academy of Sciences of the United States of America*, 72(5):1909.

[234] Wong, W. S., Solomon, B. D., Bodian, D. L., Kothiyal, P., Eley, G., Huddleston, K. C., Baker, R., Thach, D. C., Iyer, R. K., Vockley, J. G., et al. (2016). New observations on maternal age effect on germline de novo mutations. *Nature Communications*, 7(1):1–10.

[235] Wright, C., Prigmore, E., Rajan, D., Handsaker, J., McRae, J., Kaplanis, J., Fitzgerald, T., FitzPatrick, D., Firth, H., and Hurles, M. (2019). Clinically-relevant postzygotic mosaicism in parents and children with developmental disorders in trio exome sequencing data. *Nature Communications*, 10(1):1–11.

[236] Wright, C. F., Fitzgerald, T. W., Jones, W. D., Clayton, S., McRae, J. F., Van Kogelenberg, M., King, D. A., Ambridge, K., Barrett, D. M., Bayzetinova, T., et al. (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *The Lancet*, 385(9975):1305–1314.

[237] Wright, C. F., FitzPatrick, D. R., and Firth, H. V. (2018). Paediatric genomics: diagnosing rare disease in children. *Nature Reviews Genetics*, 19(5):253.

[238] Wu, X., Pang, E., Lin, K., and Pei, Z.-M. (2013). Improving the measurement of semantic similarity between gene ontology terms and gene products: insights from an edge-and ic-based hybrid method. *PloS One*, 8(5).

[239] Xu, B., Roos, J. L., Levy, S., Van Rensburg, E., Gogos, J. A., and Karayiorgou, M. (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. *Nature Genetics*, 40(7):880.

[240] Yang, S., Wang, L., Huang, J., Zhang, X., Yuan, Y., Chen, J.-Q., Hurst, L. D., and Tian, D. (2015). Parent–progeny sequencing indicates higher mutation rates in heterozygotes. *Nature*, 523(7561):463–467.

[241] Yang, Y., Muzny, D. M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., et al. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA*, 312(18):1870–1879.

[242] Yauk, C. L., Berndt, M. L., Williams, A., Rowan-Carroll, A., Douglas, G. R., and Stämpfli, M. R. (2007). Mainstream tobacco smoke causes paternal germ-line DNA mutation. *Cancer Research*, 67(11):5103–5106.

[243] Yilmaz, R., Beleza-Meireles, A., Price, S., Oliveira, R., Kubisch, C., Clayton-Smith, J., Szakszon, K., and Borck, G. (2015). A recurrent synonymous KAT6B mutation causes say-barber-biesecker/young-simpson syndrome by inducing aberrant splicing. *American Journal of Medical Genetics Part A*, 167(12):3006–3010.

[244] Young, L. C., Hartig, N., del Río, I. B., Sari, S., Ringham-Terry, B., Wainwright, J. R., Jones, G. G., McCormick, F., and Rodriguez-Viciana, P. (2018). SHOC2–MRAS–PP1 complex positively regulates RAF activity and contributes to noonan syndrome pathogenesis. *Proceedings of the National Academy of Sciences*, 115(45):E10576–E10585.

[245] Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics*, 10:451–481.

[246] Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934.